Heikki Alanen

Massadatan analytiikka ja datan hyödyntäminen yrityksessä

Metropolia Ammattikorkeakoulu

Insinööri (AMK)

Mediatekniikan koulutusohjelma

Insinöörityö

29.4.2015

Metropolia

Tekijä Otsikko Sivumäärä Aika	Heikki Alanen Massadatan analytiikka ja datan hyödyntäminen yrityksessä 48 sivua + 1 liite 29.4.2015
Tutkinto	Insinööri (AMK)
Koulutusohjelma	Mediatekniikka
Suuntautumisvaihtoehto	Digitaalinen media
Ohjaaja	Yliopettaja Kari Aaltonen

Insinöörityön tarkoituksena oli selvittää, mitä big data on ja mistä tässä paljon palstatilaa saaneessa ja keskustelua herättäneessä ilmiössä on kyse. Lisäksi tarkoituksena oli luoda dokumentoitu prosessi, jonka avulla suurten tietomassojen käyttö onnistuu pk-yrityksissä. Luotua prosessia ja työkaluja oli tarkoitus havainnollistaa kaupallisella datan analysointiin tarkoitetulla pilvipalvelulla.

Prosessikuvauksessa pyrittiin huomioimaan kaikki oleelliset työvaiheet, joita tarvitaan datan käyttöönotossa ja toimijat, jotka ottavat osaa prosessin suoritukseen. Prosessi kuvattiin kaaviona ja sanallisena esityksenä.

Big datan käyttöönottoprosessia ja tarvittavia työkaluja havainnollistettiin Hadoop-ohjelmistokehykseen pohjautuvan kaupallisen pilvipalvelun avulla. Demonstraatiossa käytettiin ohjelmankehitysympäristöä tarvittavien ohjelmien kirjoittamiseen ja kääntämiseen. Lisäksi käytettiin komentotulkkia pilvipalvelun etäkäyttämiseksi.

Prosessin luonnin ja demonstraation tekemisen aikana havaittiin, että prosessin suorittamiseksi vaaditaan asiantuntija tai ryhmä, jolla on kokemusta liiketoiminta-alueelta ja laaja-alainen koulutus tietojenkäsittelyn, ohjelmoinnin, matematiikan ja tilastotieteen alueilta. Lisäksi henkilöltä tai ryhmältä vaaditaan taitoa visualisoida prosessin tulokset ja kommunikoida ne kohdeyleisölle.

Avainsanat	big data, suuret tietomassat, Hadoop, MapReduce, HDInsight

Author Title Number of Pages Date	Heikki Alanen Big data analytics and utilization of the data in a company 48 pages + 1 appendix 29 April 2015
Degree	Bachelor of Engineering
Degree Programme	Media Technology
Specialisation option	Digital Media
Instructor	Kari Aaltonen, Principal Lecturer

The purpose of this study was to analyze the characteristics of big data and why this phenomenon has caused so much discussion. In addition, the aim was to create a documented process that allows large data masses to be used successfully in small and medium-sized firms. The created process and tools were to be demonstrated with a commercial cloud computing service.

The target of the process description was to take into account all relevant events and players who take part in the execution of the process. The process was described as a diagram that is discussed in detail in the thesis.

The implementation process of big data and the necessary tools were demonstrated with a commercial cloud computing service. The service is based on the Apache Hadoop framework, with which it is fully compatible. An integrated development environment was used in the demonstration to create the needed demonstration programs in the local development environment. In addition, a command line tool was used to control cloud services remotely.

It was found that there is a need for an expert or a group of experts, who has experience of a related business area and education in data processing, programming, mathematics and statistics. In addition, the person or the group is required of skills to visualize the results of the process and to communicate them to the target audience.

Keywords	big data, Hadoop, MapReduce, HDInsight

Sisällys

Liitteet

Liite 1. Hype Cycle for Emerging Technologies, 2014

Lyhenteet

API	Application Programming Interface. Ohjelmointirajapinta.
CAPTCHA	"**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part". Kuvavarmennus.
DVD	Digital Video Disc. Optinen datan tallennusväline.
HDFS	Hadoop Distributed File System. Hajautettu ja skaalattava tiedostojärjestelmä Hadoop-ohjelmistokehykseen.
IaaS	Infrastructure as a Service. Infrastruktuuri palveluna, malli jossa laiteympäristö on ulkoistettu ja käyttäjä hallinnoi ympäristöä haluamallaan tavalla.
MIT	Massachusetts Institute of Technology. Massachusettsin teknillinen korkeakoulu.
PaaS	Platform as a Service. Palvelinalusta palveluna, malli jossa käyttäjä voi suorittaa koodaamiaan ohjelmia pilvipalvelussa.
reCAPTCHA	Kuten CAPTCHA, mutta lisättynä kuvalla, jonka sisällön selvittämiseen vaaditaan ihmistä.
SaaS	Software as a Service. Verkkosovelluspalvelu, malli jossa ohjelmisto hankitaan palveluna ja siitä maksetaan käytön mukaan.
SDK	Software Development Kit. Joukko työkaluja, jotka mahdollistavat ohjelmien kehityksen tietylle ympäristölle.
UTF-8	Unicode-merkistöstandardin koodaustapa.

1 Johdanto

Insinöörityön tarkoituksena on selvittää, mitä big data on, kuinka dataa kerätään ja miten sitä voidaan käsitellä, analysoida ja visualisoida. Lisäksi selvitetään datan merkitystä liiketoiminnalle ja sitä, millaisia ovat suurten tietomassojen käsittelyyn tarkoitetun ympäristön arkkitehtuuri ja työkalut yleisellä tasolla. Työn toisena päätarkoituksena on myös kehittää prosessi, joka helpottaisi datan käyttöönottoa yrityksissä. Työssä tarkastellaan myös Microsoftin pilvipalvelun tarjoamia työkaluja datan analysointiin. Näiden työkalujen käyttöä ja luotua prosessia havainnollistetaan esimerkin avulla.

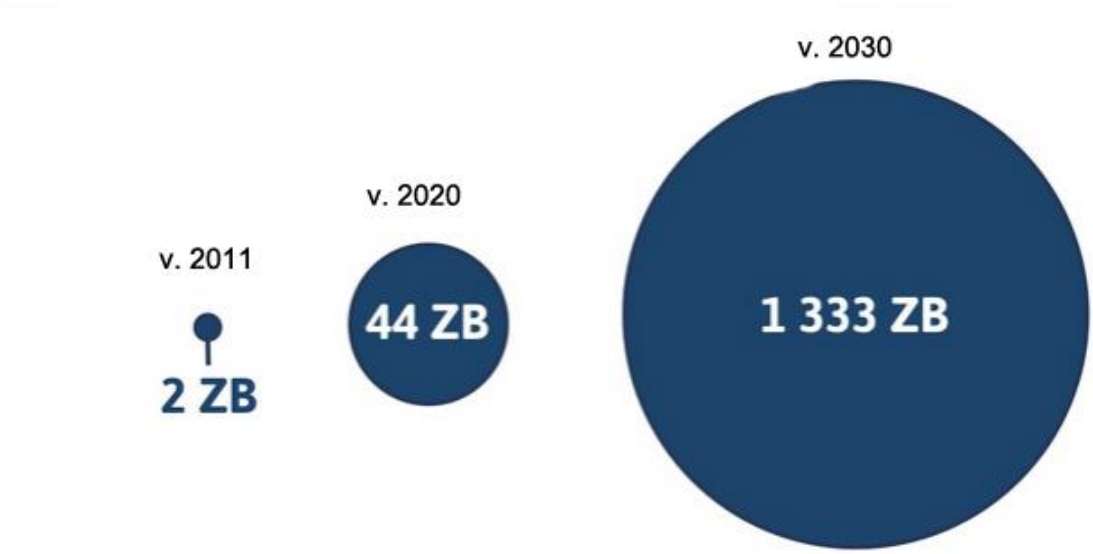
Big data, suomeksi massadata tai iso data, on paljon käytetty termi. Koska aihealueen suomenkielinen termistö ei ole vakiintunutta ja koska kotimainenkin ammattikirjallisuus käyttää pääasiassa englanninkielistä termiä big data, englanninkielistä termiä käytetään myös tässä insinöörityössä. Big data -termiä käytetään kuvaamaan sellaisia vaihtelevia tietoja tai suuria tietomääriä, joiden käsittely perinteisillä menetelmillä olisi hankalaa, jollei jopa mahdotonta. Tyypillisesti ne ovat luonteeltaan järjestelemätöntä, sisällöltään vaihtelevaa ja jatkuvasti lisääntyvää dataa. Esimerkiksi käyvät vaikkapa tuotantoprosessien mittausdata, maksukorttien käyttötiedot, sosiaalisen median tiedot, elektroniset kirjat tai laitteiden paikannustiedot. Listaa voisi jatkaa loputtomiin, sillä nyky-yhteiskunnassa tietoa kerätään jatkuvasti enemmän.

Aikaisemmin kerättyä dataa analysoitiin tarkkojen tietoaineistosta otettujen näytteiden avulla, koska käytetyt tekniikat ja menetelmät eivät tukeneet koko tietomäärän käyttöä aineistona. Vasta viimeisten vuosien aikana teknologia on kehittynyt sellaiselle tasolle, että suurten tietomassojen varastointi ja analysointi kokonaisuutena on tullut mahdolliseksi. Koko tallennetun aineiston käyttäminen analysoinnissa on myös avannut uusia mahdollisuuksia datan käytölle. Aineistoa voidaan käyttää pääasiallisen käyttötarkoituksensa lisäksi myös muuhun tarkoitukseen, jolla ei välttämättä ole lainkaan yhteyttä alkuperäiseen käyttötarkoitukseensa.

Big data on ollut ja on edelleen paljon käytetty termi ja muotisana tietojenkäsittelyssä. Sen luomiin mahdollisuuksiin on kohdistunut jopa ylisuuria odotuksia. Viimeisimmällä julkaistulla Gartnerin Hype Cycle for Emerging Technologies 2014 -käyrällä se on ohittanut käyrän lakipisteen, joten ennusteen mukaan on odotettavissa pettymyksiä, koska ylisuuret odotukset eivät täyttyneetkään [1, liite 1].

2 Big data

Tietojenkäsittelyn maailma on muuttumassa varsin merkittävästi, datan määrä kasvaa ja samalla sen käyttömuodot muuttuvat. Kuva 1 havainnollistaa arvioitua datan määrän kasvua maailmassa. Tiedon määrä on jo nyt niin valtava, että sitä on vaikea hahmottaa. Yksi zettatavu on 10^9 teratavua ja 10^{21} tavua. Yhteen teratavuun mahtuu noin 217 DVD-levyn verran dataa. Yksi zettatavullinen dataa tallennettuna DVD-levyille muodostaisi pinottuna noin 282 000 km korkean tornin. [2, s. 6.]

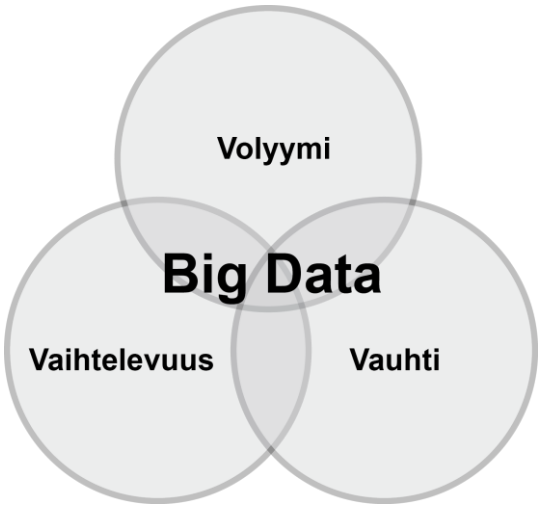


Kuva 1. Datan arvioitu määrä maailmassa [3].

Kehityssuunta aiheuttaa suuria haasteita tietojenkäsittelyssä, mutta luo samalla uusia liiketoimintamahdollisuuksia.

2.1 Big data käsitteenä

Big data -käsiteellä ei ole olemassa yhtä virallista määritelmää, vaan se on yleiskäsite, jota yleisimmin kuvataan kuvan 2 esittämillä kolmella ominaisuudella: volyymi, vaihtelevuus ja vauhti (englanniksi three V's: Volume, Variety and Velocity). Nämä kolme ominaisuutta esiteltiin jo vuonna 2001, kun verkkokauppatoiminnan kasvaessa oli havaittu datamäärien kasvavan huomattavasti. Tämän lisäksi dataa luotiin nopeasti ja vaihtelevissa muodoissa. [2, s. 26; 4.]



Kuva 2. Datan kolme V:tä [5, s. 23].

Vaikka näiden kolmen ominaisuuden esittelyn yhteydessä ei vielä puhuttu big data -käsitteestä, oli sen peruselementit tunnistettu. Tämä tapahtui noin kymmenen vuotta ennen kuin big data ilmiönä alkoi saada julkisuutta. Näitä peruselementtejä käytetään kirjallisuudessa yhä edelleen.

Volyymi

Volyymilla tarkoitetaan datan suurta määrää. Uuden teknologian avulla pystytään dataa luomaan, keräämään ja tallentamaan yhä tehokkaammin. Älypuhelimissa on 20 megapikselin kameroita, joiden tuottamat valokuvat voidaan siirtää automaattisesti pilvipalveluihin. Sosiaalisen median palvelujen viestejä ja päivityksiä voidaan kirjoittaa ja julkaista helposti älypuhelimilla. Lisäksi niissä on erilaisia sensoreita, jotka voivat tallentaa tietoja samalla kun älypuhelinta käytetään. Muun muassa paikkatieto voidaan liittää kuviin, sosiaalisen median palveluihin ja muihin sovelluksiin. Usein big data -käsite ymmärretään vain siten, että kyseessä on suuret tietomäärät. Volyymilla voidaan kuitenkin tarkoittaa myös tietueiden, tapahtumien, taulukoiden tai tiedostojen määrää [6, s. 6]. On myös huomattava, että volyymi on vain yksi osa big datan käsitettä. [5, s. 10–11.]

Vaihtelevuus

Big datalle on tyypillistä datan vaihtelevuus ja se, että dataa voi tulla hyvinkin erilaisista lähteistä. Data ei myöskään ole aina selkeästi strukturoitua, eli sillä ei ole tarkkaa rakennetta, joka mahdollistaisi sen tallentamisen esimerkiksi perinteisiin relaatiotietokan-

toihin. Esimerkiksi kirjat muodostavat sellaisen datalähteen, jota ei voi esittää tauluina ja riveinä tietokannassa. Data voi olla strukturoitua tai täysin strukturoimatonta tai sitten jotakin näiden kahden ääripään väliltä, jolloin puhutaan semistrukturoidusta datasta. Esimerkkinä strukturoidusta datasta voidaan ajatella kaupan asiakastietoja, joissa ovat yhteystiedot ja asiointihistoria. Videokuva on esimerkki strukturoimattomasta datasta. Metatiedoilla, kuten milloin ja missä video on kuvattu ja keitä videolla esiintyy, voidaan videokuvasta saada aikaan semistrukturoitua dataa. Nykyään suurin osa tuotetusta datasta on strukturoimatonta, ja sen osuus vain kasvaa, joten tarve uusille datan käsittelymenetelmille on suuri. Tätä tarvetta yrittää tyydyttää big data uusine menetelmineen. [2, s. 27; 5, s. 25.]

Vauhti

Kolmas olennaisesti big data -käsitteeseen liittyvä asia on vauhti tai nopeus, jolla dataa tuotetaan, kerätään ja käsitellään. Varsinkin erilaiset tuotantoprosessit, verkkopalvelimet, ympäristöä seuraavat sensorit ja muut laitteet tuottavat jatkuvasti uutta dataa, joko tasaisesti tai ryöppyinä. Tämä tuotettu data pitäisi myös pystyä tallentamaan nopeasti analysointia varten. Toisaalta vauhti viittaa siihen, että tuotettuun dataan pitäisi pystyä reagoimaan nopeasti. Tällöin voidaan käsitellä datavirtaa heti ja tallentaa vasta jälkeenpäin. [2, s. 27.]

Näiden vuonna 2001 esiteltyjen peruselementtien pohjalta ovat useat tahot kehittäneet omia big data -määritelmiään. Esimerkiksi tutkimus- ja konsultointiyritys Gartnerin määritelmä koostuu nykyään kolmesta osasta, jonka ensimmäinen osa on jo edellä mainittu kolmen V:n malli. Lisäksi tulevat teknologiat datan käsittelyyn kustannustehokkaasti ja innovatiivisesti sekä kolmantena ja tärkeimpänä lisääntyneet näkemykset ja parantunut päätöksenteko. Ohjelmistovalmistaja SAP Business Innovation taas lisää kolmen V:n lisäksi vielä kaksi uutta ominaisuutta. Ensimmäisenä on arvo (Value), jolla tarkoitetaan datan arvoa yrityksen liiketoiminnalle. Dataa hyödyntämällä voidaan löytää monia tapoja kustannusten alentamiseen, mutta suurin hyöty sen käytöstä tulee sellaisista käyttötavoista, jotka lisäävät myyntituloja. Toinen lisätty ominaisuus on datan laatu ja ymmärrettävyys (Veracity). [7; 8.]

Edellä esitellyt kolme alkuperäistä ominaisuutta mainitaan puhuttaessa big datasta, mutta eroaako big data perinteisestä datasta jollakin tapaa? Käytännössä ei mitään suuria eroja ole; rajanveto näiden välille on vaikeaa tai jopa mahdotonta. Big datan

sijasta voitaisiin myös puhua suurista, nopeasti kasvavista ja sisällöltään monimuotoisista datamassoista. Yksi erityinen alue kuitenkin liittyy big data -käsitteeseen, avoin data. Sillä tarkoitetaan esimerkiksi julkishallinnon avaamia tietokantoja tai muiden toimijoiden avaamia datavarantoja, joita voi käyttää maksutta tai maksua vastaan. On arvioitu, että potentiaali näissä avoimissa datavarannoissa on jopa suurempi kuin organisaatioiden omissa datavarannoissa. [2, s. 35.]

2.2 Ihminen osana ilmiötä

Nykyään yksittäinen ihminen toimii, ehkä tietämättäänkin, osana big data -ilmiötä. Verkossa toimija jättää aina jälkensä eri yritysten ja organisaatioiden sivuille ja tietoja operaattoreille ja palveluntarjoajille. Sivuilla käynnistä syntyy dataa, vaikkei mitään ostaisikaan. Yritys saa tietoa sivuilla käyvistä ihmisistä, mahdollisista asiakkaista, ja heidän kiinnostuksenkohteistaan. Usein tämä kerätty data myös jaetaan jonkun kolmannen osapuolen kanssa erilaisia analyyseja varten. Fyysisestä kaupassa käynnistäkin jää dataa yritysten käyttöön. Kaupan kävijämääriä voidaan seurata, kanta-asiakaskortin avulla voidaan henkilöidä ostotapahtuma ja liittää tapahtumaan ostokset, käytetyt maksutavat, käytetty rahamäärä ja tallentaa nämä kaikki asiakasrekisteriin. [5, s. 40.]

Kuluttajan näkökulmasta katsottuna voi suurten tietomassojen käyttö olla joko uhka tai mahdollisuus. Datasta saatava hyöty riippuu paljon siitä, mitä sen analysoija on halunnut saada aikaan. Suurimpina uhkina voidaan pitää yksityisyyden menetystä ja tietosuojan heikkenemistä. Näitä aiheita käsitellään tarkemmin jäljempänä. Positiivisia vaikutuksia datan käytöllä voi olla esimerkiksi kaupankäynnissä. Tarkentuneet tiedot kuluttajan ostokäyttäytymisestä ja tarpeista sekä yrityksen omien toimintatapojen pullonkauloista mahdollistavat henkilökohtaisen tarjonnan ja palvelun sekä tehostuneiden toimintatapojen kautta edullisemmat tuotteet. [5, s. 40–41.]

Kuluttajalla on periaatteessa mahdollisuus välttää tietojensa luovuttaminen, mutta käytännössä se voi olla hankalaa. Kanta-asiakaskortteja ei välttämättä tarvita, myöskään ei ole välttämätöntä ostaa tavaroita tai palveluja verkkokaupoista, eikä markkinointitutkimuksiinkaan tarvitse välttämättä osallistua. Tosin silloin ostokset voivat tulla kalliimmiksi ja asiointi kaupoissa voi olla hankalampaa. Kanta-asiakaskorteilla saa kohdennettuja tarjouksia tai muita hyötyjä korvauksena datansa luovuttamisesta. Lisäksi pank-

ki- ja kanta-asiakaskortteihin liitetyt lähimaksutekniikat helpottavat kassalla asiointia. [5, s. 44.]

2.3 Uhka yksityisyydelle

Samaan aikaan, kun datan avoimuuteen kannustetaan jopa Euroopan unionin taholta [8], on herännyt huoli tietoturvasta ja yksityisyydestä. Erityisesti julkisen sektorin avatessa tietovarantojaan on vielä epäselvää, millainen vaikutus uusilla datan käytön mahdollisuuksilla voi olla yksityisyydelle. Tehokkaat menetelmät kerätä ja tallentaa dataa yhdessä tehokkaiden analysointimenetelmien kanssa saattavat lisätä myös erilaisia väärinkäytöksiä ja uhata yksittäisen ihmisen yksityisyyttä. Voidaan esimerkiksi kuvitella tilanne, jossa kaupassa oleva kameravalvontajärjestelmä seuraa asiakasvirtoja ja kasvontunnistusalgoritmien avulla voidaan yksittäiset asiakkaat erottaa asiakasvirrasta. Jos lisäksi saadaan käyttöön kuvia sosiaalisesta mediasta, henkilö voidaan tunnistaa ja hänelle voitaisiin tarjota kaupassa yksilöllistä palvelua. Suomessa henkilörekistereitä koskevat lait ovat varsin tiukkoja, joten esimerkki ei ole laillisesti mahdollinen, mutta väärinkäytösten mahdollisuus on olemassa tekniikan kehittyessä. [5, s. 40.]

Massachusettsin teknillisen korkeakoulun (MIT) tutkijat tekivät tutkimuksen, jossa käytettiin yli miljoonan ihmisen luottokorttiostosten tietoja kolmen kuukauden ajalta. Luottokortteja oli käytetty 10 000 liikkeessä. Tutkimuksen tarkoituksena oli selvittää, voiko datasta, jossa ei ole henkilöiden tunnistetietoja, paljastaa käyttäjän henkilöllisyyden. Datasta oli puhdistettu kaikki muu tieto pois paitsi ostokseen käytetty rahamäärä, päivämäärä, liikkeen tyyppi ja koodi jokaiselle käyttäjälle. Tutkimuksessa käytettiin toisena tietolähteenä simuloitua aika- ja paikkatietodataa, joka liitettiin jokaiseen käyttäjään. Käytännössä ihmiset tuottavat itsestään samanlaista dataa joka päivä. Esimerkiksi valokuvissa on aikaleima ja usein myös paikkatieto. Lisäksi älypuhelimissa on sovelluksia, jotka käyttävät paikkatietoa ja niin edelleen. Kun nämä kaksi data-aineistoa yhdistettiin ja etsittiin korrelaatiota tietojen välillä, tutkijat huomasivat, että jo neljällä paikkatiedolla pystyttiin käyttäjistä 90 % selvittämään. Vaikka alkuperäiset tiedot olivat ilman yksilöintitietoja, lisäämällä tietoja toisesta lähteestä saatiin käyttäjien henkilöllisyys selville. [10.]

Toisaalta on myös tapauksia, joissa erilaisten data-aineistojen avulla on saatettu jopa pelastaa ihmishenkiä. Suomessa tällainen tapaus sattui vuonna 2011, kun yksi ihminen

menehtyi syötyään myrkyllisiä oliiveja. Kuolemantapauksen jälkeen elinturvallisuusvirasto Evira määräsi vialliset tuotteet vedettäväksi markkinoilta [11]. Tuotteen myyjällä ei ollut suoraan tietoa asiakkaista, jotka olivat tuotetta ostaneet. Kanta-asiakasohjelman rekisteriin ei tuolloin kerätty tuotekohtaista tietoa asiakkaan ostoksista vaan ainoastaan tuoteryhmäkohtaista tietoa. Jotta vaarallisen tuotteen ostajat saatiin selville, piti yhdistää kahden tietojärjestelmän sisältämät tiedot. Kassajärjestelmästä saatiin selville ostettu viallinen tuote sekä viite kanta-asiakastietoihin ja kanta-asiakasjärjestelmästä tuotteen ostaneen asiakkaan tiedot. Näin saatiin asiakkaita varoitettua ja estettyä lisävahingot. Suomessa tällainen henkilötietojen etsiminen eri tietokantojen tietoja yhdistelemällä vaatii viranomaisen määräyksen ja luvan. [12.]

Suomessa henkilötietojen käsittelyä säädellään henkilötietolailla. Lain tarkoituksena on toteuttaa yksityiselämän suojaa ja muita yksityisyyden suojaa turvaavia perusoikeuksia henkilötietoja käsiteltäessä ja edistää hyvän tietojenkäsittelytavan kehittämistä ja noudattamista [13]. Liikenne- ja viestintäministeriö asetti vuoden 2013 lopussa työryhmän, jonka tehtävänä oli luoda kokonaiskuva big datasta Suomessa ja laatia luonnos kansalliseksi big data -strategiaksi [14]. Tietosuojavaltuutettu vastasi lausuntopyyntöön ja kommentoi datan käyttöä näin:

> Jos tiedot anonymisoidaan, voidaan tietoa hyödyntää vapaammin. Anonymisoinnin on kuitenkin oltava tehokasta siten, ettei tietoa voida enää palauttaa henkilötiedoksi. [15.]

Aiemmin esitellyn Massachusettsin teknillisen korkeakoulun tutkimuksen mukaan datan anonymisointi ei välttämättä taannut henkilöiden yksityisyydensuojaa, vaan tietoaineistoja yhdistelemällä voitiin yksittäiset kuluttajat tunnistaa. Tämä voikin aiheuttaa ongelman datan anonymisoinnissa: mikä on se määrä tietoa, mikä täytyy jättää pois, jotta täytetään henkilötietolain vaatimukset, mutta ei kuitenkaan menetetä datan käyttöarvoa. MIT:n tutkimus osoitti, että sinänsä turvalliselta näyttävä aineisto voidaan avata.

2.4 Big data ja tietoturva

Kuten monen muunkin raaka-aineen kohdalla, pelkän datan arvo ei ole välttämättä suuri. Sen arvo syntyy, kun sitä käsitellään, yhdistellään ja analysoidaan ja saadaan aikaan tuloksia, jotka voivat olla yritykselle hyvinkin arvokkaita. Toisaalta jos raakadata joutuu vääriin käsiin, siitä voi aiheutua oikealle omistajalle merkittäviä haittoja esimer-

kiksi julkisuuskuvan heikkenemisenä ja sitä kautta suoranaisina tappioina liiketoiminnassa. Samoin voi käydä, jos vanhentunut eli näennäisesti käyttökelvoton raakadata päätyy asiattomien haltuun. Raakadata ei välttämättä ole tietovarkauksien kohde, vaan pääasiallinen kohde voi olla jo käsitellyt tiedot ja tulokset, jotka voivat olla hyvinkin arvokkaita. Lisäksi datan käsittelyyn käytettävät algoritmit voivat olla kohteena tietovarkauksille. Tietomurtojen taustalla voi myös olla tarkoitus väärentää raakadataa siten, että datan analysoinnin tulos ohjautuu haluttuun suuntaan, joka sitten hyödyttää rikollista tahoa. [2, s. 50–52.]

Koska tekninen kehitys on hyvin nopeaa, tulee lainsäädäntö esimerkiksi tietosuojan osalta hitaasti perässä. Lakien säätämisvaiheessa on mahdotonta ennakoida uusia tulevia tekniikoita, ja prosessi itsessään on hidas. Toisaalta on myös niin, ettei olemassa olevia lakeja noudateta. Esimerkiksi useissa verkkopalveluissa rekisteritietojen keräämisestä tai evästeiden käytöstä ei kerrota käyttäjälle mitenkään, vaikka näin lakien ja säädösten mukaan pitäisi toimia. Oman vaikeutensa aiheuttaa se, että verkkopalveluiden liiketoiminta voi olla maailmanlaajuista, jolloin on mahdotonta tutkia jokaisen maan lainsäädäntöä erikseen. [2, s. 55; 5, s. 44.]

Tietoturvan osalta ongelmia aiheuttaa suurten datamäärien tallennus- ja laskenta-arkkitehtuurin rakenne, jollaiseen esimerkiksi suosittu big data -ratkaisu Hadoop pohjautuu. Näitä järjestelmiä ei alun perin ole suunniteltu siten, että niillä käsiteltäisiin salassa pidettäviä aineistoja. Sama koskee muita avoimen lähdekoodin työkaluhankkeita, joita käytetään aineistojen käsittelyyn ja tallennukseen. Näitä kaikkia on vaikea suojata perinteisillä tavoilla. Palvelimet ja muut laitteet voidaan suojata perinteiseen tapaan, mutta suojausmenetelmät aiheuttavat hitautta datan käsittelyyn. Tätä ei datan nopeaan käsittelyyn suunnitelluissa järjestelmissä haluta. [2, s. 52.]

Pilvipalveluiden käyttö datan käsittelyssä on yleistymässä niiden tarjoamien helposti laajennettavien resurssien vuoksi. Toisaalta näiden palvelujen tarjoajiin kohdistuu vielä epäilyksiä luotettavuuden suhteen. Tietoturvan osalta eivät ainakaan suuret toimijat halua vaarantaa liiketoimintaansa. Yhdenkin asiakkaan tietojen vuotaminen voisi aiheuttaa yritykselle suuren vahingon. Tältä osin yritysten on turvallista käyttää pilvipalveluja omien data-aineistojensa analysointiympäristönä. [2, s. 53.]

3 Datan kerääminen ja käsittely

Dataa syntyy hyvin paljon, hyvin erilaisilla alueilla. Nämä alueet voidaan jakaa karkeasti neljään eri osaan: digitaalinen todellisuus, fyysinen todellisuus, saatavilla olevat datavarannot ja potentiaalinen data. Digitaalinen todellisuus käsittää esimerkiksi internetympäristön ja sosiaalisen median. Fyysinen todellisuus kattaa kaikenlaiset sensorit, jotka tuottavat dataa, ja myös niin sanotun esineiden internetin. Saatavat datavarannot koostuvat avoimesta datasta ja datamarkkinoista. Potentiaalinen data tarkoittaa dataa jota, ei vielä kerätä mutta jolla voisi tulevaisuudessa olla merkitystä. [3.]

3.1 Verkkopalvelut ja sosiaalinen media

Hyvin usein yritykset keräävät dataa vastikkeellisesti, mutta rahan sijaan antavat vastikkeeksi jotain muuta. Markkinointitutkimuksessa kysytään kuluttajalta erilaisia yritystä kiinnostavia asioita ja vastikkeeksi tarjotaan arvontavoitto [5, s. 41]. Tällä tavalla saadaan kuluttaja kiinnostumaan tutkimuksesta ja myös vastaamaan siihen. Ilman palkintoa vastausprosentti voisi olla pienempi. Toinen tapa saada kuluttajan tietoja on tarjota vastikkeeksi palveluja. Yksi tällainen tietoja keräävä palvelu on verkossa toimiva Balancion [16]. Palvelu tarjoaa käyttäjälle oman talouden hallintaan toimintoja, joilla voi seurata omaa rahankäyttöään, laatia budjetteja ja säästösuunnitelmia ja seurata niitä. Jotta palvelu olisi mahdollista toteuttaa, datalähteeksi tarvitaan käyttäjän tilitiedot, jotka saadaan pankkien tarjoamien rajapintojen avulla. Palvelun tarjoava yritys siis kerää käyttäjien tilitietoja ja saa näin tarkkaa dataa käyttäjien rahankäytöstä. Palvelun käyttöehtojen mukaan käyttäjä antaa suostumuksensa käyttää anonyymeiksi muutettuja tietojaan yrityksen tarpeisiin sekä suostumuksensa luovuttaa näitä tietoja myös kolmansille osapuolille [17].

Facebook toimii tiedonkeräyksessään samanlaiseen tapaan kuin edellä esitelty Balancion, mutta huomattavasti suuremmassa mittakaavassa. Rekisteröitymällä palveluun käyttäjä luovuttaa palvelun tuottavalle yritykselle tietonsa sekä oikeuden luomaansa datavirtaan. Maksuksi käyttäjä saa palvelun, joka tuottaa hänelle lisäarvoa. Saamaansa dataa Facebook käyttää apuna profiloidessaan käyttäjiä. Profiloinnin tuloksena syntyneitä kuluttajasegmenttejä myydään mainostajille, jolloin saadusta datasta syntyy liiketoimintaa. Mainostajat taas saavuttavat saatujen tietojen avulla juuri oikeat kohteet mainoksilleen. [5, s. 42.]

Internetin palvelujen käyttäjiä käytetään myös apuna datan digitalisoimisessa. Verkkopalvelujen kehittäjillä on käytössään työkalu nimeltä CAPTCHA, jolla varmennetaan, että palvelun käyttäjä on ihminen eikä automaattinen tietokoneohjelma. Sopivassa tilanteessa, esimerkiksi rekisteröitymisen yhteydessä, käyttäjälle näytetään kuva, jossa on kirjaimia tai numeroita hieman epäselvästi esitettynä. Käyttäjän tehtävänä on kirjoittaa kirjaimet tai numerot viereiseen kenttään, ja palvelu saa varmistuksen, että käyttäjä on todella ihminen. Google tarjoaa verkkopalvelujen kehittäjille jatkokehitettyä versiota nimeltä reCAPTCHA. Siinä käyttäjälle näytetään kaksi kuvaa. Ensimmäistä kuvaa käytetään ihmisen tunnistamiseen kuten CAPTCHA-palvelussa, toinen esitettävä kuva on sellainen, jonka sisältöä Googlen automaattinen tekstintunnistusohjelma ei ole pystynyt selvittämään. Kun käyttäjiltä on saatu tarpeeksi monta yhtenevää vastausta, kuvan sisältö hyväksytään osaksi Googlen tietokantaa. [18, s. 98–99.]

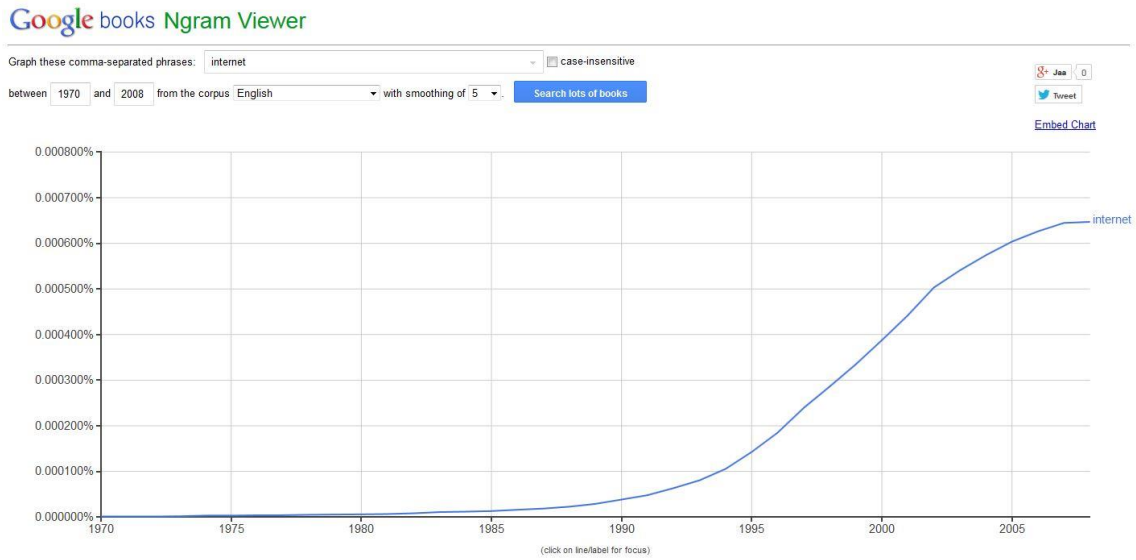
3.2 Olemassa olevan datan digitalisoiminen

Vuonna 2004 Google aloitti varsin mittavan työn. Tavoitteena oli digitalisoida kaikki julkaistut kirjat ja antaa internetin käyttäjille mahdollisuus käyttää niitä ilmaiseksi. Yhtiö ryhtyi yhteistyöhön maailman suurimpien ja arvostetuimpien akateemisten kirjastojen kanssa ja skannasi miljoonia kirjoja. Jotta tämä olisi ollut taloudellisesti mahdollista, yhtiö kehitti laitteen, joka pystyi automaattisesti kääntämään kirjojen sivuja. Skannauksen tuloksena Googlella oli tallennettuna kirjojen sivut korkearesoluutioisina kuvina. [18, s. 83.]

Tämä ei kuitenkaan vielä täyttänyt Googlen alkuperäistä ajatusta elektronisesta kirjastosta, sillä tiedon hakeminen kuvista ei ollut mahdollista. Kuvien sisältö piti saada purettua sellaiseen muotoon, että sitä pystyi käsittelemään ohjelmallisesti. Tähän työhön käytettiin erityistä tekstintunnistusohjelmaa, jolla kuvasta saatiin muodostettua tekstimuotoista dataa. Nyt kirjojen informaatio ei ollut enää vain ihmisten luettavissa joko kirjoina tai kuvina, vaan myös tietokoneohjelmat pystyivät käyttämään muodostettua dataa hyväksi. [18, s. 83–84.]

Tämä Googlen elektroninen kirjasto on yksi esimerkki siitä, miten aikaisemmin tietojärjestelmien ulottumattomissa oleva data voidaan digitalisoida ja muuttaa sellaiseen muotoon, että sitä voidaan hyödyntää ohjelmallisesti erilaisten ongelmien ratkaise-

miseksi. Googlen elektronista kirjastoa hyödyntää internetissä toimiva Google books Ngram Viewer -sovellus.



Kuva 3. Google books Ngram Viewer [19].

Palvelulla voidaan esimerkiksi selvittää hakusanan tai hakusanojen esiintymistiheys kirjallisuudessa vuosittain. Esimerkissä on selvitetty, miten usein sana ”internet” esiintyy kirjallisuudessa vuosina 1970–2008. Tuloksen graafinen esitys on kuvassa 3.

3.3 Ajattelutavan muutos datan käsittelyssä

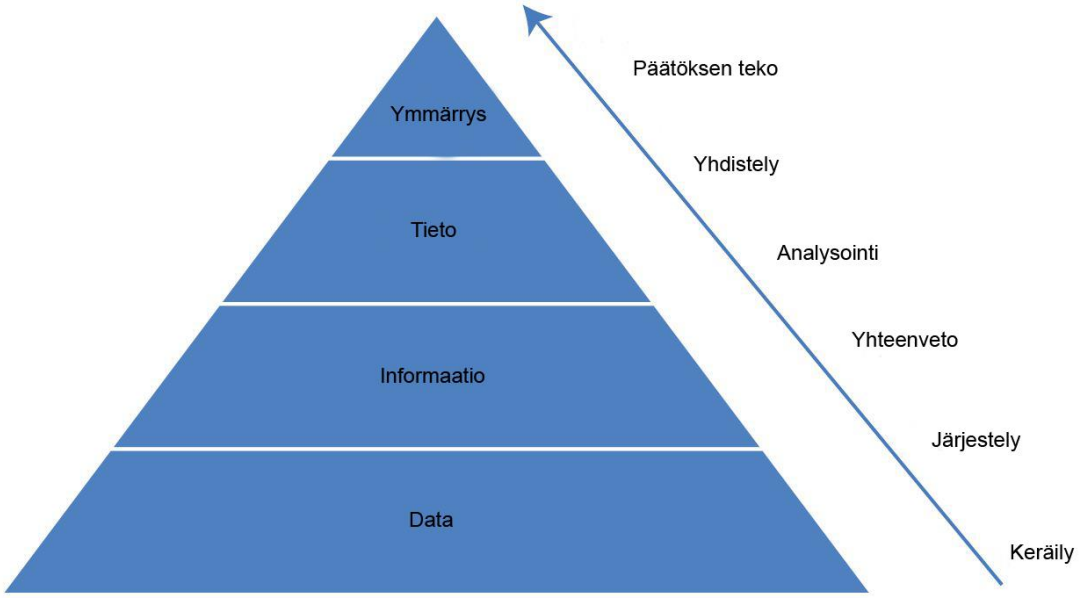
Nykyään on mahdollisuus käyttää analysoinnissa koko kerättyä data-aineistoa, mutta sillä on myös hintansa. Aineistossa on käytännössä aina epätarkkuuksia, sillä datamäärän kasvattaminen kasvattaa myös virheellisten arvojen määrää. Perinteisesti aineisto on ollut vain pieni näyte koko data-aineistosta. Tällöin näytteen arvojen tarkkuus on ollut olennaista. Mahdolliset epätarkkuuksien aiheuttajat pyrittiin poistamaan jo dataa kerättäessä, ja itse näyteaineistosta puhdistettiin sinne tulleet täysin virheelliset arvot. Big data -ajatusmaailmassa yksi suurimmista muutoksista on oppia hyväksymään kaikki ne epätarkkuudet ja virheet, joita koko kerätyn data-aineiston käytöstä aiheutuu. Hyötynä koko data-aineiston käytössä on, että siitä voi löytyä sellaisia yksityiskohtia, joita ei löytyisi käytettäessä ”puhdistettuja” näytteitä. [18, s. 32.]

Jotta dataa voidaan käyttää tehokkaasti hyväksi, se vaatii käyttäjiltä kolme suurta ajattelutapojen muutosta, jotka liittyvät kaikki toisiinsa ja samalla vahvistavat toisiaan. Ensimmäinen muutos on se, että nyt meillä on mahdollisuus analysoida suuria määriä dataa sen sijaan, että olisimme pakotettuja käyttämään pienempiä näytteitä. Toinen on se, että hyväksytään datan mahdollinen sekalaisuus ja virheellisyys sen sijaan, että se olisi aina oikeaa ja virheetöntä. Kolmas muutos ajattelutavoissa on korostaa korrelaation merkitystä. Korrelaation avulla data-analyysissa selvitetään, mitä tapahtuu tai on tapahtumassa, sen sijaan että etsittäisiin vaikeasti löydettävää syysuhdetta. [18, s. 19.]

Big datan käyttöön liittyy siis tietynlainen sekasotku. Tämä voi aiheutua ensinnäkin siitä, että virheiden todennäköisyys kasvaa, kun mittauspisteitä tai mittaustaajuutta lisätään. Esimerkkinä lämpötilan mittaaminen viinitarhassa: Yhdellä lämpömittarilla mitattaessa tulee mittarin olla tarkka ja luotettava. Sen tulee toimia jatkuvasti, eikä mittausvirheitä saa tulla. Jos lämpömittareita sijoitetaan useita eri puolille viinitarhaa, voidaan käyttää halvempia ja hieman yksinkertaisempia mittareita. Tällöin saadaan enemmän dataa, ehkä silloin tällöin vääriäkin arvoja, mutta suuremmasta määrästä dataa saadaan parempi kokonaiskuva viinitarhan lämpötilasta. Jos taas mittaustaajuutta nostetaan, saadaan dataa kerättyä enemmän, mutta arvojen aikajärjestys saattaa sekoittua. Kuitenkin suuremmalla datamäärällä voidaan saavuttaa suurempi hyöty kuin mikä on se haitta, jonka pieni sekasotku arvoissa ja aikajärjestyksessä aiheuttaa. Toisaalta sekasotkua lisää myös se, että on mahdollista yhdistää hyvinkin erilaisia datalähteitä, jotka eivät ole yhteneviä muodoltaan. [18, s. 33–34.]

3.4 Datan analysointi

Pelkällä datalla ei useinkaan ole minkäänlaista arvoa, mutta kun sitä käsitellään, yhdistellään ja analysoidaan, siitä voi muodostua käyttökelpoisia ja arvokkaita tuloksia. Kuvassa 4 on esitetty yksi näkemys, miten raakadata jalostuu eri vaiheiden kautta ymmärrykseksi, jonka pohjalta voidaan tehdä päätöksiä.



Kuva 4. Datan jalostuminen päätöksenteon välineeksi [21, s. 2].

Pyramidin pohjalla oleva data on ainoastaan merkkijonoja, bittejä ilman mitään erikoisempaa merkitystä. Datan voidaan sanoa olevan informaation ja tiedon raaka-ainetta. Dataa eli merkkijonoja ja bittejä järjestelemällä saadaan aikaan informaatiota, joka jo sisältää merkityksen tai tulkinnan. Informaatiosta tulee analysoinnin tuloksena tietoa, ja tällöin informaatio on ymmärretty ja omaksuttu. Pyramidin huipulla on ymmärrys, jolloin uusi opittu tieto yhdistetään aiempiin tietoihin ja kokemuksiin. Pyramidista voidaan havaita, että mitä korkeammalle tasolle edetään, sitä enemmän tarvitaan inhimillistä ajattelua, työstämistä ja arviointia. [20, s. 31.]

Tilastotieteessä käytetty käsite korrelaatio on hyödyllinen datan analysoinnissa, mutta big datan yhteydessä käytettynä korrelaatiosta saadaan esille suurin hyöty. Yksinkertaistettuna korrelaatio kuvaa kahden muuttujan välistä riippuvuutta. Vahvaksi korrelaatioksi kutsutaan tilannetta, jossa toisen muuttujan arvon muuttuessa muuttuu hyvin todennäköisesti myös toisen muuttujan arvo. Heikko korrelaatio merkitsee, ettei muuttujan arvon muutos vaikuta lainkaan toiseen tai ainakaan se ei vaikuta merkittävästi. Big datan tapauksessa korrelaatiolla voidaan vastata kysymykseen mitä, mutta datan avulla ei voida selvittää miksi jokin asia tapahtuu. Syy-seuraussuhde tulee selvittää muilla menetelmillä. [18, s. 52–53.]

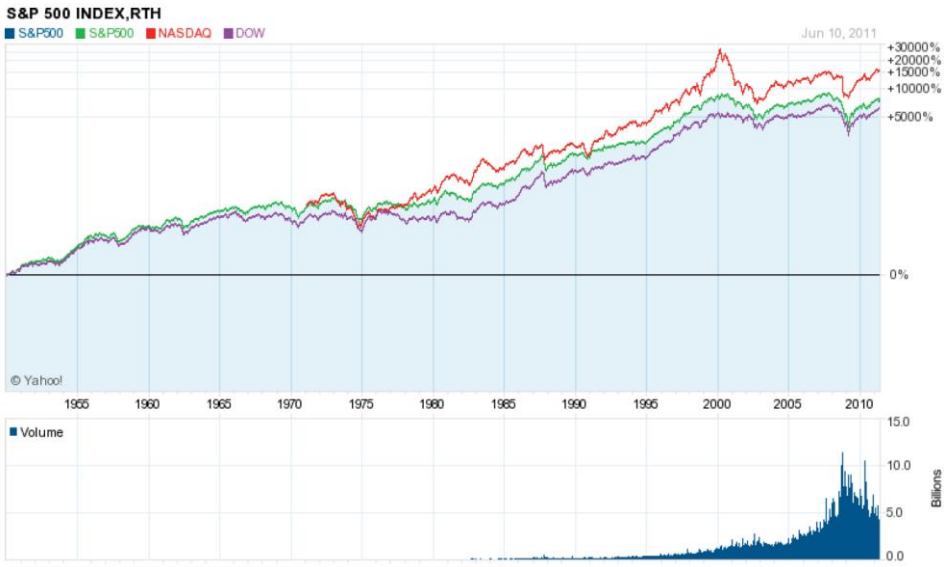
Vuonna 2009 Google käytti tallentamiaan tietoja hyväksi ennustaessaan influenssan etenemistä Yhdysvalloissa. Yhtiö tutki käyttäjien hakukoneeseen syöttämiä hakuehtoja

ja etsi niistä viittauksia influenssaan, sen hoitoon, lääkkeisiin ja niin edelleen. Oman datan lisäksi Googlella oli käytössään Yhdysvaltojen viranomaisten raportit influenssoista edellisiltä vuosilta. Näitä raportteja ja hakutuloksia vertaamalla eli hakemalla korrelaatiota erilaisten mallien avulla Google pystyi kertomaan lähes reaaliajassa, missä alueilla Yhdysvalloissa influenssa oli leviämässä. Virallisten lääkärien antamien raporttien avulla tieto olisi saatu vasta useita viikkoja myöhemmin. [18, s. 1–2.]

3.5 Tulosten visualisointi

Kaikkialla ympärillämme on erilaista infografiikkaa ja datavisualisointeja eli erilaisten tietojen visuaalisia esityksiä. Niitä voivat olla kaaviot, kartat, kuvakkeet, merkit, julisteet ja piirrokset, mutta ne kaikki eivät ole infografiikoita. Infografiikka ja datavisualisointi voidaan helposti ymmärtää synonyymeiksi, mutta graafisen alan ammattilaiset määrittelevät ne eri asioiksi. [22, s. 1–2.]

Datavisualisoinnit ovat graafisia esityksiä, joissa numeeriset arvot kuvataan visuaalisesti. Yksi esimerkki datavisualisoinnista ovat graafiset kaaviot, joissa annetusta datasta muodostetaan kuva. Graafisesta kaaviosta voidaan helposti nähdä esimerkiksi kehityssuunta ja voidaan helposti tehdä asioiden välisiä vertailuja. [22, s. 2.] Kuva 5 on esimerkki graafisesta esityksestä, jossa on kuvattu kolmen osakeindeksin kehitys 1950-luvulta lähtien.

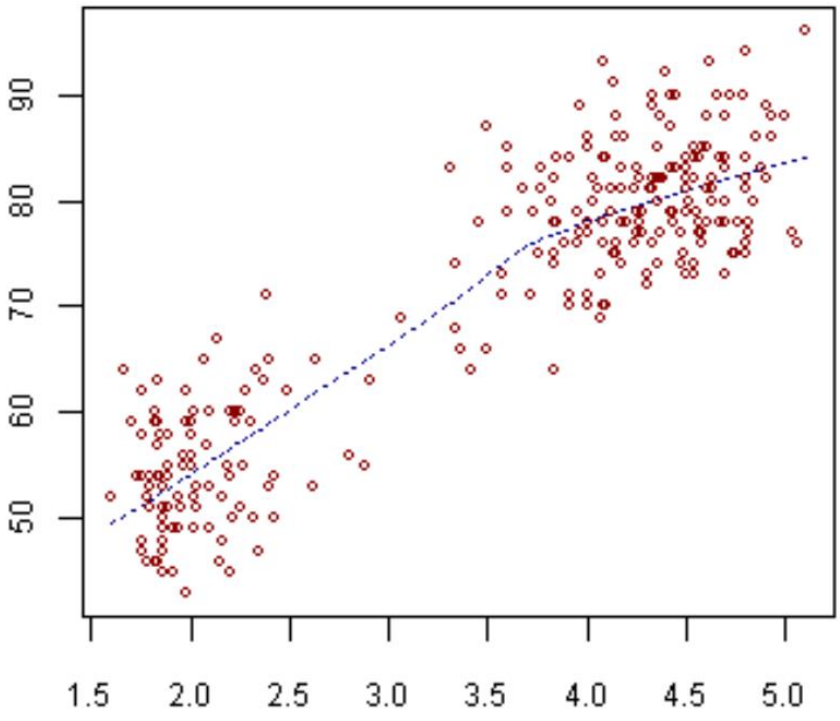


Kuva 5. Graafinen esitys osakeindeksien kehityksestä [22, s. 4].

Viivakaaviosta voidaan helposti nähdä jatkuva nouseva kehityssuunta, eri indeksien välinen suhde toisiinsa sekä pörssin huiput ja romahdukset. Kaikki tämä voidaan nähdä yhdellä sivulla, tietokoneen ruudulla tai kalvolla. Jos tämä kaikki data esitettäisiin taulukkona, olisi näitä asioita mahdotonta nähdä yhdellä silmäyksellä. [22, s. 4.]

Esityksiä muodostettaessa on myös huomattava, etteivät niiden lukijat halua nähdä kaikkea dataa yhdessä paikassa, vaan tarvitaan monitasoisia esityksiä. Esitys voidaan jakaa esimerkiksi kolmeen tasoon: graafinen yhteenveto, moniulotteinen näkymä, jolla tarkennetaan ylätason esitystä uusilla tiedoilla, ja yksityiskohtainen näkymä, jossa esitetään tiedot yksityiskohtaisesti. [21, s. 271–272.]

Erilaisia graafisia esitystapoja on paljon, ja niistä täytyy löytää oikea tapa esittämään käsiteltäviä tuloksia. Datan visualisoinnissa eräs käyttökelpoisista kaavioista on pistekaavio, josta on esimerkki kuvassa 6.



Kuva 6. Esimerkki pistekaaviosta.

Pistekaaviolla voidaan erityisesti havainnollistaa muuttujien välistä riippuvuutta eli korrelaatiota. Samaan kaavioon voidaan lisätä myös muita kuvaajia. [21, s. 275.]

Kuvan 7 esitystapa on nimeltään sanapilvi. Sitä voidaan käyttää tekstimuotoisen datan visualisointiin.

Kuva 7. Esimerkki sanapilvestä.

Sanapilven avulla voidaan havainnollistaa sanojen esiintymistiheyttä annetussa tekstissä. Kaaviossa eniten käytetty sana esitetään suurimmalla kirjainkoolla, jonka koko pienenee esiintymistiheyden pienentyessä. Myös värejä voidaan käyttää esittämään tätä asiaa. [21, s. 278.]

Kun datavisualisointiin yhdistetään kuvituksia, tekstiä ja kuvia saadaan tulokseksi infografiikkaa. Infografiikan katsotaan alan asiantuntijoiden keskuudessa olevan laajempi graafinen työ, joka parhaimmillaan kertoo lukijalle kokonaisen tarinan. Niistä on tullut yhä enemmän artikkelien tai puheiden kaltaisia, tarkoituksena joko tiedottaa, viihdyttää tai vakuuttaa. Hyvä infografiikka koostuu kolmesta osasta: johdanto, asia ja johtopäätös [22, s. 6, 27.]

4 Datan merkitys liiketoiminnalle

Näyttäisi siltä, että big data ei olisi enää pelkkä markkinointitermi. Organisaatiot eri aloilla ovat etsimässä uusia keinoja tehdäkseen parempia liiketoimintapäätöksiä. Yritykset etsivät tapoja kuinka luoda uusia tuotteita nopeammin, kuinka tuoda uusia tuotteita markkinoille ja kuinka parantaa asiakaspalvelua. Ne ovat ymmärtäneet tarpeen laajentaa käyttämiään tiedonhallintajärjestelmiään ja ottaa käyttöön uusia menetelmiä. [21, s. 6.]

4.1 Big datan potentiaali

Yhdysvaltalainen konsulttitoimisto McKinsey & Company on vuonna 2011 tutkinut, millainen potentiaali suurten tietomassojen käytöllä on eri toimialoilla. Tutkimuksessa on arvioitu eri alojen datan määrää, kuinka nopeasti dataa muodostuu, kuinka vaihtelevaa data on ja kuinka paljon käyttämätöntä dataa alalla on. Taulukossa 1 on esitetty yhteenveto tutkimuksen tuloksista.

Taulukko 1. Datan potentiaali eri aloilla [21, s. 10].

	Volyymi	**Vauhti**	**Vaihtelevuus**	**Käyttämättömän datan määrä**	**Datan potentiaali**
Pankkiala ja arvo-paperimarkkinat	suuri	suuri	pieni	keskimääräinen	suuri
Tiedonvälitys	suuri	suuri	suuri	keskimääräinen	suuri
Koulutus	hyvin pieni	hyvin pieni	hyvin pieni	suuri	keskimääräinen
Hallinto	suuri	keskimääräinen	suuri	suuri	suuri
Terveydenhuolto	keskimääräinen	suuri	keskimääräinen	keskimääräinen	suuri
Vakuutusala	keskimääräinen	keskimääräinen	keskimääräinen	keskimääräinen	keskimääräinen
Tuotanto	suuri	suuri	suuri	suuri	suuri
Luonnonvarat	suuri	suuri	suuri	suuri	keskimääräinen
Vähittäiskauppa	suuri	suuri	suuri	suuri	suuri
Kuljetusala	keskimääräinen	keskimääräinen	keskimääräinen	suuri	keskimääräinen
Tarvikkeet	keskimääräinen	keskimääräinen	keskimääräinen	keskimääräinen	keskimääräinen

Taulukosta havaitaan, että toimialan potentiaalin on katsottu olevan suuri, jos ainakin yksi big datan perusominaisuuksista, volyymi, vauhti tai vaihtelevuus, on suuri ja käyttämätöntä dataa on keskimääräisesti. Aloja, joilla odotusarvo on tutkimuksen mukaan suuri, ovat finanssiala, tiedonvälitys, terveydenhuolto, vähittäiskauppa, tuotanto ja hallinto. [21, s. 9–10.]

4.2 Datan hyödyntäminen

Asiakkailta kerätty data, joka on vain keräävän yrityksen hallussa, voi tuottaa yritykselle suurta kilpailuetua. On mahdollista, että tulevaisuudessa data on ainoa asia, joka tuottaa yritykselle kilpailuedun. Yhdistämällä yrityksen hallussa olevan data, saatavilla oleva julkinen avoin data ja maksulliset datalähteet, voidaan luoda kokonaiskuva yrityksen toiminnasta. Datan keräämisellä asiakasrajapinnasta ja sen tehokkaalla analysoinnilla voidaan esimerkiksi parantaa tuotekehitystä. Tämän lisäksi voidaan saada aikaan ratkaisuja, jotka parantavat kannattavuutta ja tehokkuutta. [5, s. 33–34.]

Monet menestyksekkäät dataa hyödyntäneet palvelut ovat lähteneet liikkeelle hyvästä ideasta, mutta idean keksijöillä on myös ollut laaja-alaiset taidot toteuttaa ideansa. Heillä ei myöskään ole välttämättä ollut omia datavarantoja, vaan palvelun ajatuksena on ollut hyödyntää jo olemassa olevia aineistoja ja niitä käsittelemällä tarjota käyttäjille uusia palveluita. Yritykset, joilla on omia datavarantoja mutta joissa ei ole taitoja hyödyntää dataa omaan käyttöön, voivat esimerkiksi lisensoida dataa muiden käyttöön ja saada näin aikaan liiketoimintaa keräämällään datalla. Toinen tapa hyödyntää omia datavarantojaan on käyttää apuna yrityksiä, jotka tarjoavat konsultointia, teknologiaa tai analytiikkapalveluja. [18, s. 124–125, 131.]

> Big datan vaikutus liiketoimintaan on mullistava. Datasta tulee keskeisin kestävä kilpailuedun lähde ja sen varaan rakennetuista innovaatioista yrityksen tärkein resurssi. Ne, jotka heräävät datan tärkeyteen nopeimmin ja suurimalla intensiteetillä, tulevat parantamaan suhteellista kilpailuasemaansa lähitulevaisuudessa nopeimmin. [5, s. 135.]

4.3 Menestystarinoita

Kun etsitään menestystarinoita datan käytöstä, tulee varmasti ensimmäisenä mieleen Facebook. Se on pystynyt keräämänsä datan avulla luomaan hyvin kannattavaa liiketoimintaa ja menestymään pörssissä. Tällaisia kansainvälisiä suuryrityksiä on varmasti muitakin, mutta onko suomalaisia yrityksiä, joissa hyödynnettäisiin kerättyä suurta datamassaa? Suomessa big data ja sen käyttö koetaan ilmeisesti vielä liikesalaisuudeksi, sillä yritykset eivät ole julkisuudessa juurikaan kertoneet hankkeistaan. Seuraavassa esitellään kolmen suomalaisen yrityksen hankkeita.

Peli- ja viihdeyhtiö Rovio kerää dataa useita teratavuja joka päivä esimerkiksi pelien suorittamiseen käytettävistä palvelimista ja pelejä pelaavien päätelaitteista. Rovio haluaa peleistä kerättävällä datalla ymmärtää käyttäjän toimintaa ja tehdä kannattavaa asiakashankintaa. Lisäksi kertyvää dataa hyödynnetään pelien kehittämisessä. Peleihin pyritään löytämään yhdistelmiä, joilla saataisiin aikaan suurin arvo. Parannuksia onkin saavutettu esimerkiksi pelien liikevaihtoon ja asiakaspysyvyyteen. Myös tiedottamista on saatu kohdennettua. Jos datasta havaitaan, että pelaajat ovat peleissä jumissa, Rovio voi lähettää pelaajille vihjeen, kuinka edetä. Lisäksi uusista päivityksistä voidaan lähettää tietoa. Rovio käyttää datan varastointiin ja analytiikkaan Amazonin pilvipalveluja. [23.]

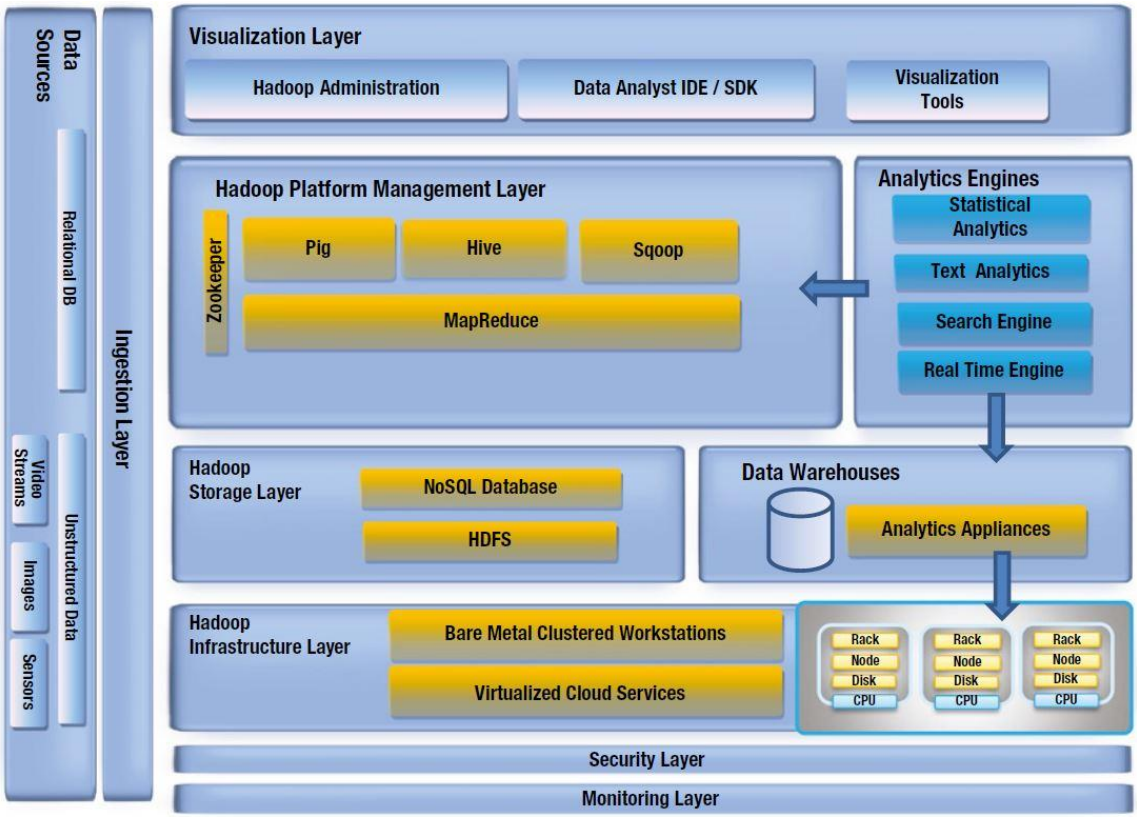
Teleoperaattori Sonera käyttää dataa apuna asiakaskokemuksen parantamisessa. Soneran tytäryhtiön Tele Finlandin käytössä on niin sanottu säästövahti-palvelu. Palvelussa seurataan asiakkaan kolmen viimeisen kuukauden laskuja, ja käytön perusteella suositellaan asiakkaan käyttöön sopivinta liittymää. Rovion tavoin myös Sonera käyttää dataa apuna parantamaan asiakaspysyvyyttä, mutta käytössä on lisäksi toisia prosesseja asiakkaista kiinni pitämiseksi. Uusin alue, jolla dataa hyödynnetään, on tekstianalytiikka. Dataa kerätään sosiaalisesta mediasta, asiakaspuheluista ja chat-palvelusta. Tästä aineistosta saadaan kuva, miksi asiakas joutuu ottamaan yhteyttä, ja löydetään asiat, jotka ovat nousussa tai laskussa. [23.]

Suomalaisista teollisuusyrityksistä ABB on ollut ensimmäisten joukossa hyödyntämässä big dataa liiketoiminnassaan. ABB kerää dataa valmistamistaan taajuusmuuttajista huollon etätukipalvelua varten. Dataa kerätään niiden mittausantureista automaattisesti Microsoft Azure -pilvipalveluun. Etätukipalvelu toimii tällä hetkellä vikatilanteita varten. Jos laitteeseen tulee ongelma, asiakas saa huollon nopeammin ja tuotantokatkokset lyhenevät. Jatkossa yrityksen on tarkoitus hyödyntää taajuusmuuttajista saamaansa dataa ennakoimalla laitteen huollon tarvetta eli sitä, onko laite huollon tarpeessa vai voisiko määräaikaishuoltoa siirtää vielä eteenpäin. Huolto pyrittäisiin ajoittamaan siten, että siitä tulisi säästöjä ja vaadittavat seisokit pystyttäisiin minimoimaan. [23.]

5 Työkalut suurten tietomassojen käsittelyyn

5.1 Big data -sovelluksen arkkitehtuuri

Ennen kuin ryhdytään rakentamaan tai ostamaan big data -palveluja, täytyy varmistaa, että kaikki tarvittavat arkkitehtuurikomponentit ovat olemassa. Oikeanlaisen järjestelmän avulla voidaan datasta saada esille arvokkaita tietoja ja saada aikaan oikeita päätelmiä. Kuva 8 esittelee yhden mahdollisen big data -arkkitehtuurimallin. Arkkitehtuurin komponentit voivat olla joko avoimen lähdekoodin ympäristöjä tai valmiiksi koottuja kaupallisia ympäristöjä. [24, s. 9.]



Kuva 8. Big data -arkkitehtuuri [24, s. 10].

Yksi arkkitehtuurin peruskomponenteista on datan lähde (Data Sources). Yrityksellä voi olla käytössään useita sisäisiä tai ulkoisia datan lähteitä, joita on pystyttävä käsittelemään nopeasti ja tehokkaasti. Ennen kuin data tallennetaan sovelluksen käyttöön, datasta suodatetaan pois selkeästi virheelliset tulokset (Ingestion Layer). Samalla dataa voidaan muokata omaan käyttöön sopivaksi. [24, s. 12–13.]

Jatkokäsittelyä varten data tallennetaan tallennuskerroksen (Hadoop Storage Layer) tiedostojärjestelmään. HDFS (Hadoop Distributed File System) on kehitetty tallentamaan hyvin suuria määriä dataa hajauttamalla se useille eri palvelimille. Fyysinen kerros (Hadoop Infrastructure Layer) toteuttaa tämän tiedon hajautuksen käytännössä. [24, s. 14–16.]

Hadoop Platform Management -kerros tarjoaa työkalut ja kyselykielet, joilla voidaan käyttää tallennettua dataa HDFS-tiedostojärjestelmässä. Tärkein osakomponentti tällä kerroksella on MapReduce. Se on ohjelmointimalli, jolla voidaan suorittaa tehokkaasti ohjelmia hajautetusti. Hadoop-ohjelmistokehykseen on kehitetty erilaisia kieliä tai työkaluja, joilla voidaan tehdä MapReduce-ohjelmia tai koordinoida eri tehtävien suoritusta. [24, s. 16–19.] Kuvassa 8 niistä on esitetty Pig, Hive, Sqoop ja Zookeeper.

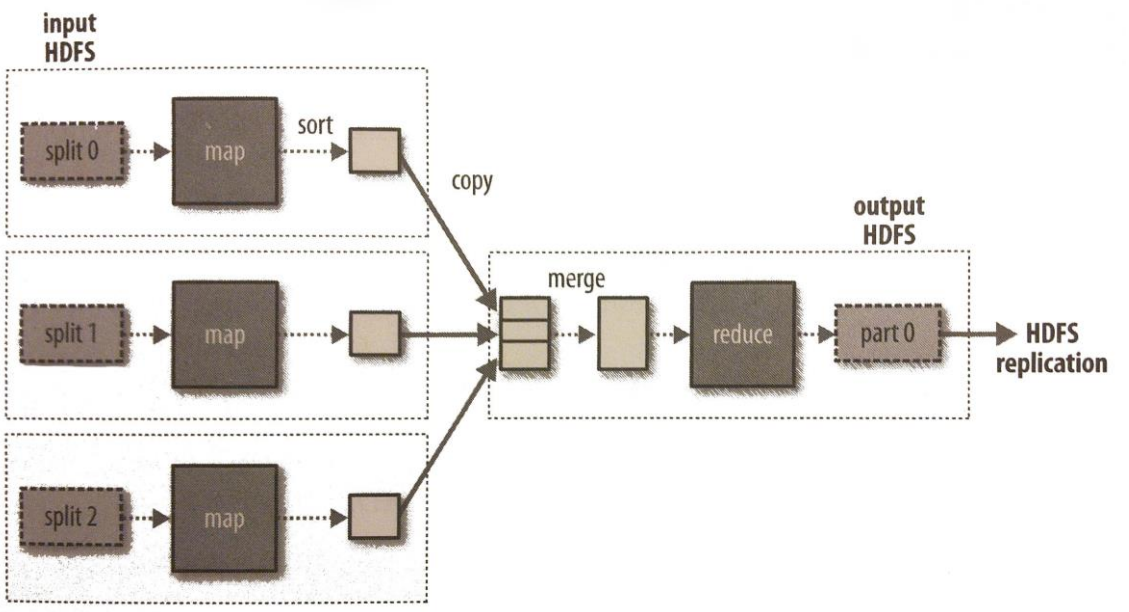
Yksi arkkitehtuurin tärkeimmistä komponenteista on tietoturva (Security Layer). Koska hyvin usein järjestelmä sisältää sellaista arkaluonteista tietoa, joka ei saa päästä vuotamaan yrityksestä ulos, tietoturvallisuusvaatimukset tulee ottaa huomioon jo järjestelmän suunnitteluvaiheessa. Toinen komponentti, joka on hyvä ottaa huomioon jo suunnittelun alkuvaiheessa, on tuloksien visualisointi (Visualization Layer). Suuri datamäärä voi johtaa tiedon ylikuormaan, mutta siitä voi olla myös apua luotaessa erilaisia näkymiä tuloksiin. Yleensä Hadoop-käsittelyn tulokset tallennetaan perinteiseen relaatiotietokantaan tai muuhun vastaavaan, jolloin tiedon analysointiin ja visualisointiin voidaan käyttää tätä tallennettua dataa. [24, s. 20, 25.]

5.2 Hadoop-ohjelmistokehys

Hadoop on käytössä suuressa osassa big data -ratkaisuja. Se on Apache Software Foundationin ylläpitämä avoimen lähdekoodin ohjelmistokehys, joka mahdollistaa suurten datamäärien hajautetun käsittelyn niin sanotuissa klustereissa. Klusterit voivat muodostua joko yhdestä tai jopa tuhansista tietokoneista, jotka kaikki tarjoavat paikallista laskenta- ja tallennuskapasiteettia. Skaalautuvuus onkin yksi Hadoopin vahvuuksista ja sen menestyksen salaisuus. Hadoop pohjautuu kahteen Googlen vuosina 2003 ja 2004 esittelemään tekniikkaan, joista ensimmäinen mahdollisti suurten datamäärien tallentamisen (Google File System) ja toinen menetelmän, kuinka näitä datamääriä voitiin analysoida nopeasti ja tehokkaasti (MapReduce). Näiden kahden innovaation pohjalta työryhmä kehitti Hadoop-ympäristön. [25; 5, s. 80.]

HDFS (Apachen nimitys tiedostojärjestelmälle) ja MapReduce ovat edelleen Hadoopin ydinprojekteja. Näiden projektien erinomaisuus perustuu ensinnäkin kustannustehokkuuteen. Klusterin tietokoneiden ei tarvitse olla supertietokoneita, vaan ne voivat olla tavallisia palvelinkoneita. Toinen seikka on hajautus. Kaikki klusterin palvelimet analysoivat dataa paikallisesti, joten sitä ei tarvitse siirrellä verkossa. [5, s. 82.]

MapReduce-työ suoritetaan kahdessa osassa: ensin map-vaihe ja sitten reduce-vaihe. Niissä suoritetaan käyttäjän koodaamia map- ja reduce-funktioita. Map-vaiheessa tehtävä hajautetaan klusterin palvelimille, jotta suuren datamäärän käsittely nopeutuisi ja kuorma tasoittuisi. Kuva 9 selvittää, kuinka data kulkee työn suorituksen aikana. Map-funktio käsittelee syötteenä saamaansa raakadataa ja tulostaa avain–arvo-pareja jatkokäsittelyä varten. Map-funktion tuottamat tulokset lajitellaan, ennen kuin ne kerätään yhteen ja tallennetaan väliaikaisesti. [26, s. 18–19.]



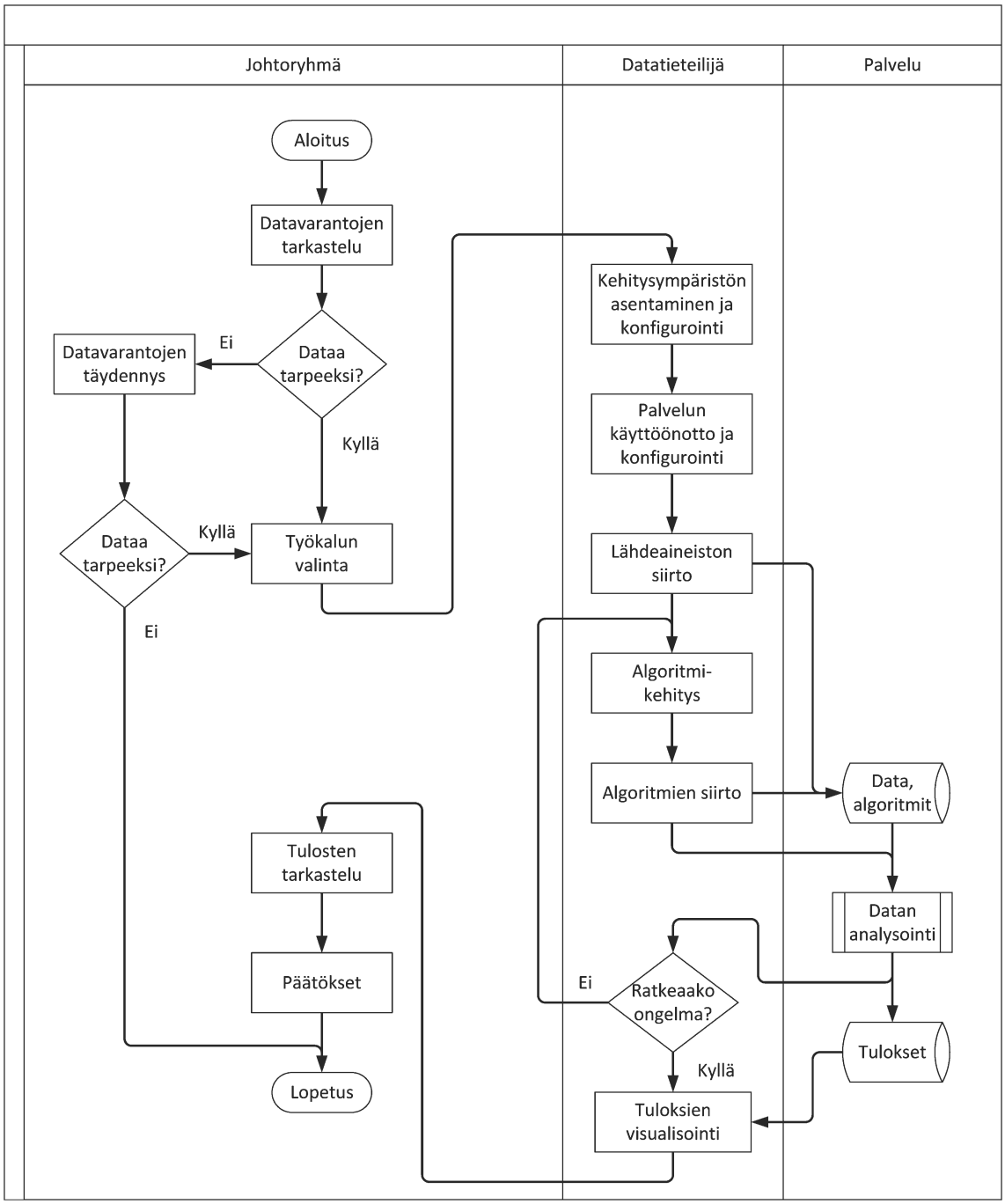
Kuva 9. MapReduce-työn datan kulku [26, s. 29].

Seuraavassa vaiheessa suoritetaan Reduce-funktio, joka lukee edellisen Map-vaiheen tulokset, muokkaa ne haluttuun muotoon ja lähettää tulokset eteenpäin tiedostojärjestelmään tallennettaviksi [26, s. 18–19, 29].

5.3 Prosessi big datan käyttöönottamiseksi

Kuten muissakin liiketoiminnan hankkeissa, myös big datan käytön aloittamiseen tarvitaan visio eli tahtotila ja suunta. Käytännössä tämä tarkoittaa sitä, että datan avulla haluttaisiin saavuttaa tuloksia, jotka tuottaisivat yritykselle esimerkiksi kustannussäästöjä, kilpailuetua tai uutta liiketoimintaa. Visiosta muodostetaan yritykselle strategia, joka kuvaa keinot ja aikataulun tahtotilan eli vision saavuttamiseksi. Prosessikuvauksella mallinnetaan, miten asetettu tavoite käytännössä saavutetaan. Prosessi on asiakkaalle arvoa tuottava tapahtumaketju, jossa asiakas voi olla joko sisäinen tai ulkoinen. Prosessi koostuu toisiinsa kytkeytyvistä tapahtumista, joilla on määritelty resurssi. [5, s. 125; 27, s. 4.]

Kuvassa 10 on esitetty mahdollinen prosessi big datan käyttöönottoon yrityksessä. Prosessikaaviossa on esitetty prosessin tapahtumat ja toimijat: johtoryhmä, datatieteilijä ja palvelu. Johtoryhmän tehtävänä on päättää prosessin aloittamisesta, millaista dataa ja kuinka paljon käytetään, mikä on työkalu datan käsittelemiseksi sekä lopuksi tulosten hyödyntäminen. Datatieteilijä on asiantuntija, joka pystyy muuttamaan liiketoimintaongelman sellaiseen muotoon, että sen ratkaiseminen on mahdollista matemaattisesti tai data-analyysin avulla. Datatieteilijän tehtäviin kuuluu myös algoritmien suunnittelu ja ohjelmointi, tuloksien visualisointi ja niiden esitteleminen päättäville tahoille. [21, s. 251–255.]



Kuva 10. Big datan käyttöönottoprosessi yrityksessä.

Big data -projektia aloitettaessa tulee selvittää, mitä kaikkea dataa on käytettävissä itsellä, yhteistyökumppaneilla tai muissa lähteissä tai miten nykyisellä järjestelmällä voisi dataa kerätä lisää [5, s. 124]. Jos dataa ei ole vielä tarpeeksi, voidaan selvittää, mistä ja miten lisädataa voitaisiin hankkia. Lisäksi on tärkeää asettaa selvä tavoite, mihin datan käytöllä pyritään.

Lähdeaineiston laadun ja määrän selvittyä voidaan miettiä, kuinka sitä käsiteltäisiin ja millaisilla välineillä. Tässä vaiheessa voidaan myös mallintaa, miten dataa tullaan käyttämään, sekä testata, ratkaiseeko malli annetun ongelman. Pieni aineistomäärä voidaan käsitellä esimerkiksi Excelillä tai muulla vastaavalla työkalulla. Mutta kun näiden yksinkertaisten välineiden suorituskyky alkaa rajoittaa datan käyttöä, tulee harkita muita vaihtoehtoja. Tietoa markkinoilla olevista vaihtoehdoista voidaan hankkia palveluja tarjoavista yrityksistä tai käyttämällä asiantuntijapalveluita tarjoavia yrityksiä eri mahdollisuuksien selvittämiseen [5, s. 124, 127].

Työkaluvalinnan jälkeen voidaan ryhtyä suunnittelemaan, asentamaan ja määrittelemään paikallista kehitysympäristöä, jos valitulla työkalulla tai palvelulla sellainen on. Muun muassa Microsoft Azure HDInsight -palvelu tarjoaa mahdollisuuden kirjoittaa ja testata ohjelmia paikallisessa kehitysympäristössä. Omana tapahtumanaan kuvan 10 prosessissa on kuvattu palvelun käyttöönotto ja konfigurointi. Esimerkiksi pilvipalvelujen käyttö vaatii palvelujen avaamisen sekä käyttäjien ja salasanojen määrittelemistä.

Lähdeaineiston siirtäminen palveluun voi aiheuttaa erityisiä toimenpiteitä, jos sitä on paljon. Nopeallakin internetyhteydellä voi aineiston siirtäminen pilvipalveluun kestää kauan. Tällöin tulee harkita vaihtoehtoisia siirtotapoja.

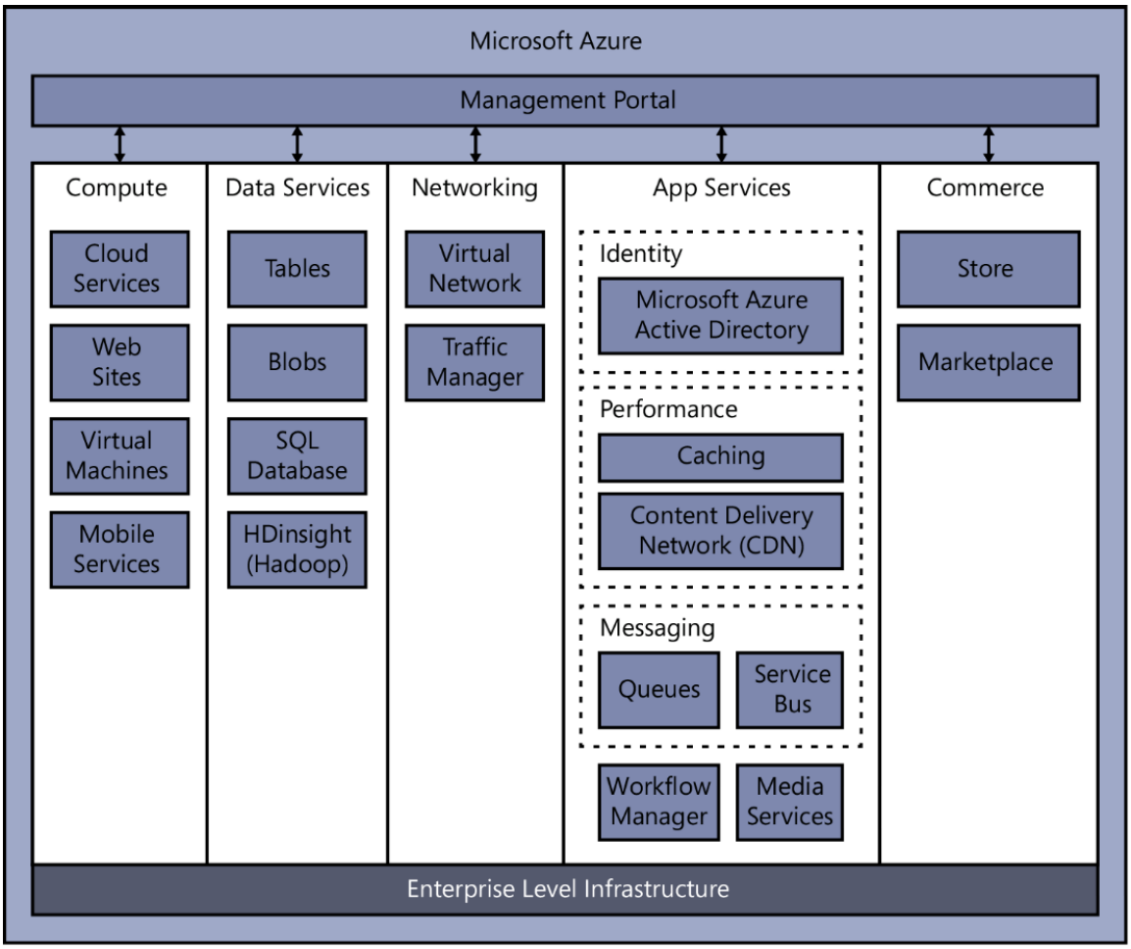
Raakadatan käsittelyyn ja analysointiin tarvittavat algoritmit voidaan kehittää käyttämällä palvelujen emulointiohjelmia paikallisessa kehitysympäristössä tai käyttämällä varsinaista pilvipalveluympäristöä kehitysympäristönä. Kun algoritmit toteuttavat halutut toiminnot ja ne on testattu, ne voidaan siirtää varsinaiseen palveluun.

Datan käsittely suoritetaan palvelussa, ja tulokset tallennetaan tiedostojärjestelmään. Tässä vaiheessa voidaan joutua tekemään iterointia algoritmien suhteen. Tulokset eivät ehkä ole täydellä aineistomäärällä sitä, mitä odotettiin, tai niistä ei pystytä tekemään mitään johtopäätöksiä, joten algoritmeihin tarvitaan muutoksia. Kun tulokset antavat vastauksen ongelmaan, datatieteilijä visualisoi tulokset helposti ymmärrettävään muotoon ja esittelee ne johtoryhmälle, minkä jälkeen johtoryhmä voi tehdä niistä päätöksiä.

6 Microsoft Azure -pilvipalvelu

6.1 Palvelut ja käyttömallit

Microsoft Azure -pilvipalvelu koostuu useista erillisistä palveluista, joita voidaan käyttää joko yksin tai yhdessä toisten kanssa. Microsoft Azure -pilvipalvelun tarjoamat palvelut on esitelty kuvassa 11. Tarjolla on esimerkiksi mahdollisuus rakentaa oma kehitysympäristö virtuaalikoneen (Virtual Machine) avulla tai rakentaa verkkosivuja ja monimutkaisia verkkopalveluja, joiden käytettävissä on erilaisia laskenta- ja tiedonvarastointipalveluja. Pilvipalvelua varten Microsoftilla on datakeskuksia, jotka mahdollistavat palvelujen tarjoamisen joko alueellisesti tai maailmanlaajuisesti. Palvelut saadaan näin mahdollisimman lähelle asiakkaita. [28, s. 7, 21–22.]



Kuva 11. Microsoft Azure -pilvipalvelu [28, s. 23].

Microsoft Azure -pilvipalvelua voidaan käyttää kolmella erilaisella tavalla. Ensimmäisessä tavassa käyttäjä kirjoittaa ja testaa oman ohjelmansa paikallisessa kehitysympäristössä (kuva 12), siirtää sen Azure-palveluun ja suorittaa kirjoittamansa ohjelman siellä. Tätä mallia kutsutaan palvelualustamalliksi (Platform as a Service, PaaS).

Toinen tapa on ulkoistaa laiteympäristö pilvipalveluun. Palvelusta otetaan käyttöön niin sanottu virtuaalikone (Virtual Machine), jonka käyttöjärjestelmäksi käyttäjä voi valita Windows- tai Linux-käyttöjärjestelmän. Ylläpitovastuu on tällöin palvelun käyttäjällä, joka voi muokata virtuaalikoneen ympäristön juuri haluamakseen. Tällaisesta mallista käytetään nimitystä palveluinfrastruktuuri (Infrastructure as a Service, IaaS).

Kolmas tapa, jolla Azure-palvelua voidaan hyödyntää, on käyttää valmiita ohjelmia, joita on tarjolla pilvipalvelussa. Tapa on nimeltään verkkosovelluspalvelu (Software as a Service, SaaS). Tässä mallissa käyttäjä ei asenna ohjelmaa paikallisesti eikä maksa kiinteää lisenssimaksua, vaan käyttää palvelua pilvipalvelussa ja maksaa käytön mukaan. Esimerkiksi Microsoft Office 365 ja HDInsight ovat tällaisia palveluja. [28, s. 23–24.]

6.2 Azure HDInsight -palvelu

Azure HDInsight on Microsoftin datan analysointiin tarjoama palvelu, joka pohjautuu Apachen Hadoop-ohjelmistokehykseen. Azure HDInsight on täysin yhteensopiva Apachen version kanssa, mikä merkitsee sitä, että kaikki Hadoopin periaatteet ja teknologiat toimivat HDInsight-ympäristössä. HDInsight-palvelun käyttöönottovaiheessa muodostetaan klusteri, jolla voi ajaa Hadoop-sovelluksia. Työn valmistuttua käyttäjä voi jättää klusterin käyntiin, pysäyttää klusterin tai tuhota sen. [28, s. 21, 26.]

Tietovarasto (Storage)

Azure HDInsight -palvelun käytön kannalta ehkä tärkein Azure-palvelun komponentti on tietovarasto lähdeaineistolle ja tuloksille. HDInsight-palvelu voi käyttää kahdentyyppisiä tietovarastoja: HDFS tai Azure Storage. Kun datan varastointiin käytetään HDFS-tietovarastoa, data on tallennettuna klusterin palvelimien tietoelementteihin ja dataa pitää käyttää ohjelmointirajapinnan kautta. Tämä merkitsee myös, että poistettaessa klusteri käytöstä häviää samalla varastoitu data.

Käyttämällä Azure Storage -palvelua tarjoutuu muutama etu verrattuna HDFS-tietovarastoon: dataa voidaan siirtää tavanomaisilla työkaluilla, poistettaessa klusteri käytöstä data säilyy varastossa, käyttö on halvempaa ja varastoituun dataan päästään käsiksi muista Azuren-palveluista ja jopa pilvipalvelun ulkopuolelta. [28, s. 25.]

Azure HDInsight -palvelu perustuu täysin Hadoop-ohjelmistokehykseen, jossa aikaisemmin esitellyt HDFS ja MapReduce ovat myös mukana. Lisäksi mukana on muita Hadoopin sisarprojekteja, jotka tukevat datan käsittelyä.

Pig

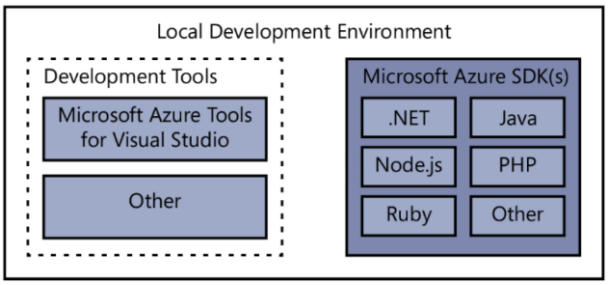
Pig on korkean tason kehitysympäristö MapReduce-funktioiden kehitykseen. Kielenä on Pig Latin, joka on tarkoitettu erityisesti suurten tietomassojen käsittelyyn ja suunniteltu käytettäväksi hajautetuissa ympäristöissä kuten Hadoop. Korkean tason kielenä sillä voidaan kirjoittaa datan käsittelyfunktioita helpommin ja nopeammin kuin esimerkiksi Javalla tai C#:lla. Se sisältää komennot datan siirtämiseen, tallentamiseen ja käsittelyyn. [28, s. 27.]

Hive

Kun halutaan työstää tietovarastossa olevaa dataa perinteisempään tapaan, voidaan käyttää ympäristöä nimeltä Hive. Ympäristöllä voidaan muodostaa tietovarastoja joko HDFS-tiedostojärjestelmän tai mahdollisten muiden tiedostojärjestelmien päälle. Käytettävä kieli on nimeltään HiveQL, joka on hyvin samantapainen kuin SQL-kieli. [28, s. 27.]

Paikallinen kehitysympäristö

Paikallisessa kehitysympäristössä voidaan ohjelmia kirjoittaa ja testata ilman, että käytettäisiin lopullista tuotantoympäristöä. Ohjelmien kehittäjän apuna käytetään Microsoft Visual Studio -ohjelmankehitysympäristöä ohjelmien kirjoittamiseen ja kääntämiseen. Kuvassa 12 on esitelty paikallisen kehitysympäristön komponentteja. Niihin kuuluu joukko ohjelmistokehitystyökaluja (SDK), joiden avulla pystytään ohjelmia luomaan eri ohjelmointikielillä ja erilaisiin käyttöympäristöihin sekä ohjelmankehitysympäristöön asennettavia lisätyökaluja (Microsoft Azure Tools).



Kuva 12. Paikallinen ympäristö Microsoft Azure -sovellusten kehittämiseen [28, s. 23].

Kuvan 12 komponenttien lisäksi tarvitaan HDInsight-emulaattori, joka jäljittelee Azure HDInsight -ympäristöä paikallisessa ympäristössä. Käyttämällä emulaattoria voidaan ohjelmat testata lähes tuotantoympäristöä vastaavassa tilanteessa. [28, s. 23, 28.]

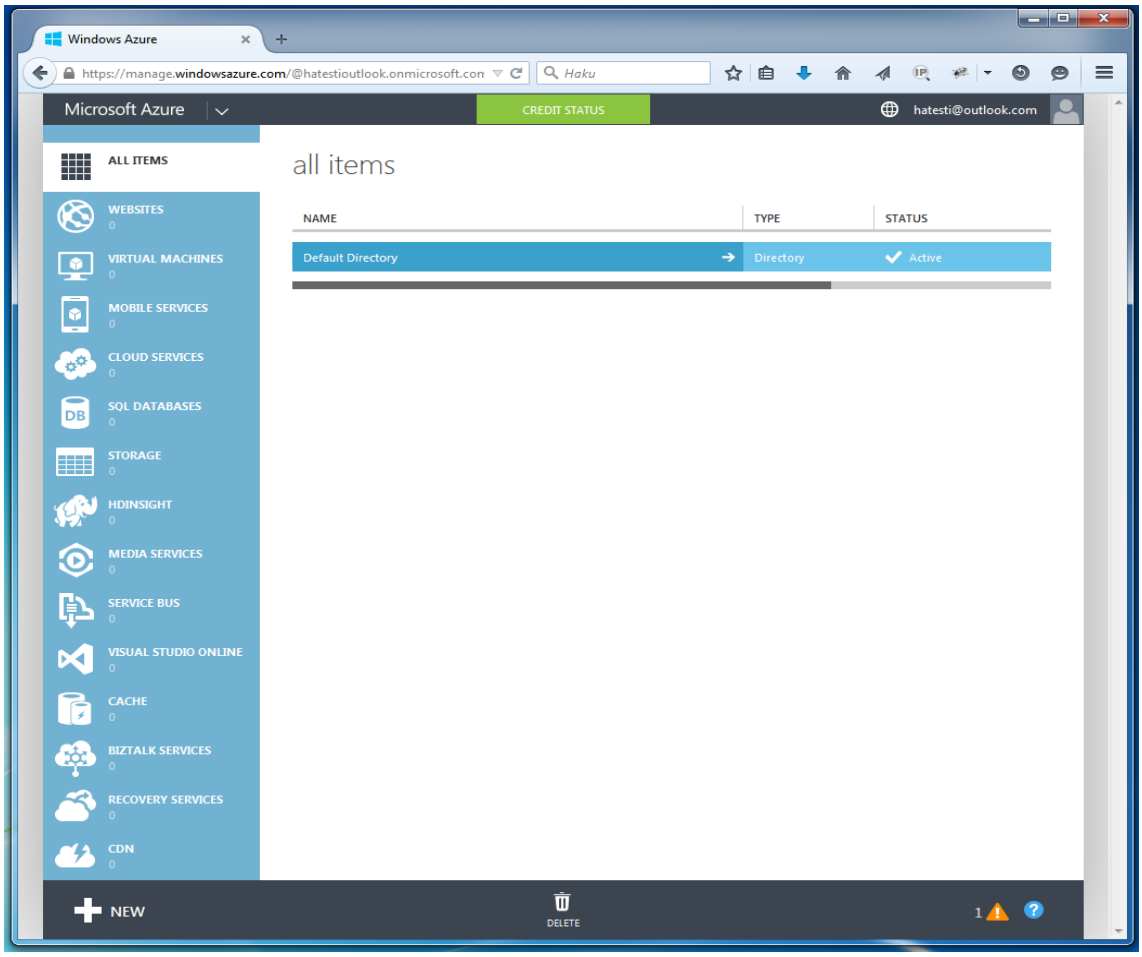
7 Datan analysointi Microsoft Azure HDInsight -palvelun avulla

Insinöörityöhön liittyvä demonstraatio suoritettiin Microsoft Azure HDInsight -ympäristössä. Tehtävänä oli laskea sanojen esiintymistiheys tekstimuotoisesta lähdeaineistosta, joka sisälsi useita tekstitiedostoja. Tarvittavien MapReduce-ohjelmien ohjelmointikielenä käytettiin C#-kieltä ja ohjelmointiympäristönä käytettiin Microsoft Visual Studio -ympäristöä.

Datavarantojen tarkastelu ja työkalun valinta

Demonstraatiossa käytettiin datalähteenä Project Gutenberg -palvelua, joka on internetissä toimiva elektroninen kirjasto [29]. Palvelussa on suuri määrä kirjoja, joiden tekijäoikeuden suoja on jo rauennut tai joiden tekijöiltä on saatu lupa materiaalin julkaisemiseen ja levittämiseen. Lähdeaineistoksi valikoitui 40 Aleksis Kiven ja Juhani Ahon kirjaa. Datan määrä ei ollut suuri, yhteensä noin 14,2 megatavua, mutta se oli riittävä demonstraation suorittamiseen.

Työkalu tehtävän suorittamiseksi oli valittu jo aikaisemmin, joten sitä ei demonstraatiossa tarvinnut tehdä. Microsoft Azure -pilvipalvelun tarjoamien palvelujen käyttämistä varten oli hankittava tilaus pilvipalveluun. Tarjolla olleista erilaisista vaihtoehdoista valittiin demonstraatiota varten niin sanottu ilmainen kokeilutilaus (Free Trial), jolla saattoi tutustua palveluun 30 päivän ajan. Microsoft Azure -pilvipalvelun suosituimmasta palvelutilauksesta maksetaan kuukausittain käytön mukaan (Pay-As-You-Go) eli vain käytetyistä resursseista maksetaan.

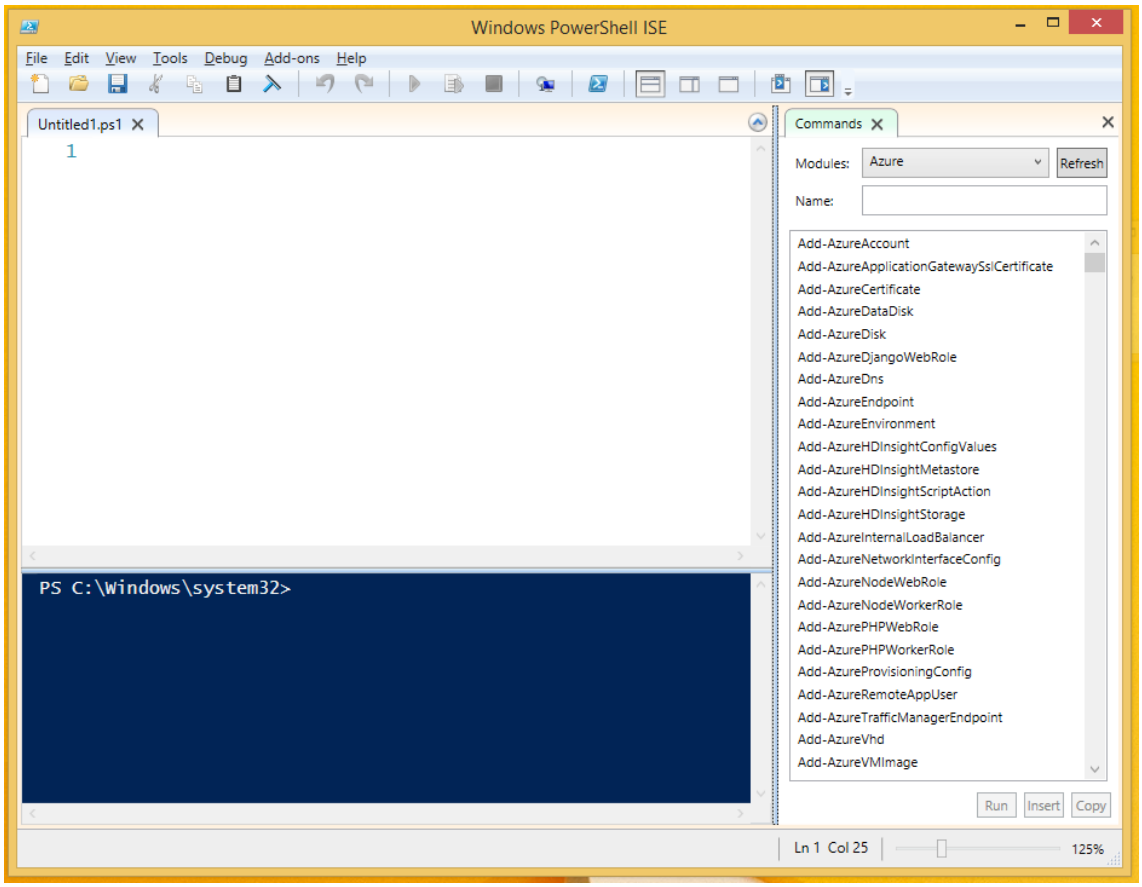


Kuva 13. Microsoft Azure -portaalin perusnäkymä.

Kuva 13 esittää näkymän Microsoft Azure -palvelun etusivulle. Tämän portaalin kautta voitiin hallinnoida tarvittavia tilauksia ja myös palveluita.

Kehitysympäristön asentaminen

Jos pilvipalvelun käyttö on toistuvaa ja samoja toimintoja toistetaan useita kertoja, kannattaa käyttöä varten kirjoittaa komentotiedostoja. Windows Powershell ISE -ympäristössä voidaan luoda, editoida, testata ja suorittaa komentotiedostoja. Perusversio Powershell-ympäristöstä tulee osana Windows-käyttöjärjestelmää. Microsoft Azure -palvelun käyttöä varten täytyy asentaa Azure Powershell -lisäosa, joka sisältää komennot pilvipalvelun käyttöön [30]. Kuvassa 14 näkyy Windows Powershell -ympäristön perusnäkymä.



Kuva 14. Windows Powershell ISE -ympäristö.

Työkaluikkunan oikean reunan komennot-näkymään (Commands) saadaan listattuna kaikki Azure-palveluun liittyvät komennot. Lisäksi editointi- ja komentorivi-näkymissä on käytössä älykäs komentojen täydennysautomatiikka, joka helpottaa komentotiedostojen luontia.

Palvelun käyttöönotto

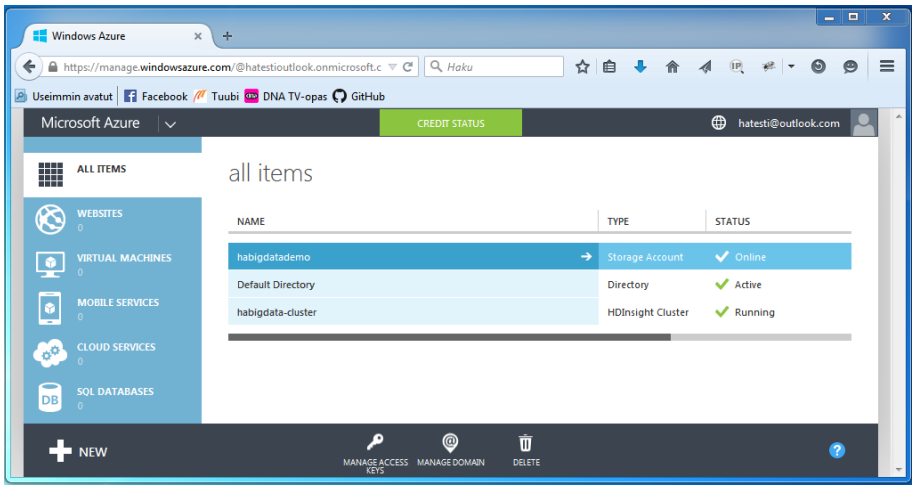
Kun paikallinen käyttöympäristö oli saatu valmiiksi, voitiin luoda ympäristö Microsoft Azure -palveluun. Kuvassa 15 on esitetty komentotiedosto, jonka riveillä 7–9 luodaan tili tietovarastolle ja riveillä 12–13 muodostetaan itse tietovarasto.

Tietovaraston luonnin jälkeen muodostettiin varsinainen HDInsight-klusteri. Kuvan 15 riveillä 36–38 on esitetty komennot klusterin luomiseen. Luonnin yhteydessä oli huomioitava, että muodostettava klusteri oli samassa datakeskuksessa kuin tietovarasto, eli tässä tapauksessa alue oli Pohjois-Eurooppa (North Europe).

```
1  # Alustetaan muuttujat
2  $subscriptionName = "Free Trial"
3  $storageAccountName = "habigdatademo"
4  $storageContainerName = "habigdatademo-container"
5
6  # Luodaan tili tietovarastolle
7  New-AzureStorageAccount -Location "North Europe" -StorageAccountName $storageAccountName
8  $storageAccountKey = (Get-AzureStorageKey -StorageAccountName $storageAccountName).Primary
9  Set-AzureSubscription -SubscriptionName $subscriptionName -CurrentStorageAccount $storageAccountName
10
11 # Luodaan tietovarasto
12 $destContext = New-AzureStorageContext -StorageAccountName $storageAccountName -StorageAccountKey $storageAccountKey
13 New-AzureStorageContainer -Name $storageContainerName -Context $destContext
14
15 # Alustetaan muuttujat HDInsight palvelulle
16 $clusterName = "habigdata-cluster"
17
18 # Alustetaan klusteri HDIsight-palvelulle. Se täytyy olla samassa datakeskuksessa kuin tietovaraston tili
19 $location = Get-AzureStorageAccount -StorageAccountName $storageAccountName | %{$_.Location}
20 $clusterNodes = 4
21
22 # Käyttäjänimi ja salasana Hadoop käyttäjälle.
23 $hadoopUserName = "hal2testi"
24 $hadoopUserPassword = "Kokeilu#21" # Salasana määritelty tässä vain demotarkoituksessa.
25 # Salasanan salaus
26 $secPassword = ConvertTo-SecureString $hadoopUserPassword -AsPlainText -Force
27 $credential = New-Object System.Management.Automation.PSCredential($hadoopUserName,$secPassword)
28
29 # Ladataan tietovaraston avaintieto
30 Select-AzureSubscription $subscriptionName
31 $storageAccountKey = Get-AzureStorageKey $storageAccountName | %{ $_.Primary }
32 $containerName = $storageContainerName
33
34 # Luodaan HDInsight klusteri
35 Write-Host "Luodaan $clusterName"
36 New-AzureHDInsightCluster -Name $clusterName -Credential $credential -Location $location `
37 -DefaultStorageAccountName $storageAccountName -DefaultStorageAccountKey $storageAccountKey `
38 -DefaultStorageContainerName $containerName -ClusterSizeInNodes $clusterNodes -ClusterType Hadoop
39
```

Kuva 15. Komennot tietovaraston ja HDInsight-palvelun luomiseksi.

Muodostettava klusteri asetettiin rivillä 38 Hadoop-tyyppiseksi -ClusterType -parametrilla. Komentotiedostossa määritellään myös HDInsight-palvelun salasana, mutta todellisessa tuotantoympäristössä salasanan tulisi olla talletettuna joko tiedostoon tai käyttöjärjestelmän rekisteriin salatussa muodossa.



Kuva 16. Tietovarasto ja HDInsight-klusteri luotuna Azure-palveluun.

Kuvassa 16 on näkymä Azure-portaaliin tietovaraston ja HDInsight-klusterin luonnin jälkeen. Näkymästä nähdään myös käyttöönotettujen palvelujen tila.

Lähdeaineiston siirto

Kun tietovarasto on pilvipalvelussa olemassa, voidaan aineisto siirtää sinne kuvan 17 komennoilla.

```
1  # Alustetaan muuttujat
2  $subscriptionName = "Free Trial"
3  $storageAccountName = "habigdatademo"
4  $storageContainerName = "habigdatademo-container"
5
6  # Tietovarasto
7  $storageAccountKey = (Get-AzureStorageKey -StorageAccountName $storageAccountName).Primary
8  Set-AzureSubscription -SubscriptionName $subscriptionName -CurrentStorageAccount $storageAccountName
9  $destContext = New-AzureStorageContext -StorageAccountName $storageAccountName -StorageAccountKey $storageAccountKey
10
11 # Siirretään datatiedostot paikalliselta levyltä Azure-palveluun
12 $localDataFolder = "C:\Users\Heikki\SkyDrive\Opinnaytetyo\demo\data"
13 $destDataFolder = "habigdata/data"
14 $files = Get-ChildItem $localDataFolder
15 foreach($file in $files) {
16     $fileName = "$localDataFolder\$file"
17     $blobName = "$destDataFolder/$file"
18     Write-Host "Copying $fileName to $blobName"
19     Set-AzureStorageBlobContent -File $fileName -Container $storageContainerName -Blob $blobName -Context $destContext -Force
20 }
21 Write-Host "Kaikki tiedostot hakemistosta $localDataFolder ladattu $storageContainerName hakemistoon"
22
```

Kuva 17. Lähdeaineiston siirto Microsoft Azure -palveluun.

Lähdeaineisto voi koostua useista tekstitiedostoista, jotka kaikki siirretään riveillä 14–20 yksitellen pilvipalveluun.

Algoritmikehitys (MapReduce-funktiot)

Tehtävän eli sanojen lukumäärän laskemiseksi tarvittiin kaksi pientä funktiota. Funktiot kirjoitettiin C#-kielellä Microsoft Visual Studio -ohjelmointiympäristössä. Hadoop MapReduce -työ suoritetaan kahdessa erillisessä vaiheessa. Ensimmäisessä vaiheessa suoritetaan WordCountMapper, jonka koodi on esitetty esimerkkikoodissa 1. Ohjelma lukee syötetiedostoa rivi kerrallaan, pilkkoo rivin yksittäisiin sanoihin ja poistaa tulkinnan kannalta ongelmalliset merkit. Lisäksi sana muutetaan pieniksi kirjaimiksi, jotta isolla alkukirjaimella alkava sana ei olisi eri sana kuin kokonaan pienillä kirjaimilla kirjoitettu sana. Lopuksi ohjelma kirjoittaa sekä sanan että arvon konsoliulostuloon. Hadoop tallentaa tulokset väliaikaiseen tiedostoon.

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.IO;
using Microsoft.Hadoop.MapReduce;
using System.Globalization;

namespace WordCountMapper {
   class WordCountMapper {
      static void Main(string[] args) {
         if (args.Length > 0) {
            Console.SetIn(new StreamReader(args[0]));
         }

         string line;
         string[] words;
         char[] charsToTrim = { ',', '.', '\"', '?', '!', '/' };
         char[] delimiters = new char[] { ' ' };

         while ((line = Console.ReadLine()) != null) {
            words = line.Split(delimiters, StringSplitOptions.RemoveEmptyEntries);

            foreach (string word in words) {
               string word2 = word.TrimStart(charsToTrim);
               string word3 = word2.TrimEnd(charsToTrim);
               Console.WriteLine(word3
                  .Replace("-", string.Empty)
                  .Replace("(", string.Empty)
                  .Replace(")", string.Empty)
                  .Replace("_", string.Empty)
                  .Replace("\"", string.Empty)
                  .Replace("\xBB", string.Empty)
                  .ToLower(new CultureInfo("fi-FI", false)) + "\t1");
            }
         }
      }
   }
}
```

Esimerkkikoodi 1. WordCountMapper.cs.

MapReducer-työn toisessa vaiheessa suoritetaan esimerkkikoodissa 2 esitelty WordCountReducer. WordCountReducer-ohjelma lukee sanojen virtaa ja laskee eri sanojen kokonaismäärän ja kirjoittaa tuloksen konsoliulostuloon. Sen jälkeen Hadoop tallentaa tulosvirran tiedostoon.

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.IO;
using Microsoft.Hadoop.MapReduce;

namespace WordCountReducer {
    class WordCountReducer {
        static void Main(string[] args) {
            string line, word, lastWord = null;
            int count = 0;

            if (args.Length > 0) {
                Console.SetIn(new StreamReader(args[0]));
            }

            while ((line = Console.ReadLine()) != null) {
                string[] split = line.Split(new Char[] { '\t' });
                word = split[0]; // key
                int value = Convert.ToInt16(split[1]); // value

                if (word != "") {
                    if (word != lastWord) {
                        if (lastWord != null) {
                            Console.WriteLine("{0}\t{1}", lastWord, count);
                        }
                        count = 1;
                        lastWord = word;
                    }
                    else {
                        count += value;
                    }
                }
            }
            Console.WriteLine("{0}\t{1}", lastWord, count);
        }
    }
}
```

Esimerkkikoodi 2. WordCountReducer.cs.

Hadoop-ohjelmistokehys lajittelee WordCountMapper-ohjelmalta tulevan tulosvirran automaattisesti, joten siitä ei reducer-funktion kehittäjän tarvitse huolehtia, kuten ei myöskään tulosten tallentamisesta tiedostoon.

Algoritmien siirto

Kuvan 18 komennoilla siirretään käännetyt MapReduce-ohjelmat palveluun.

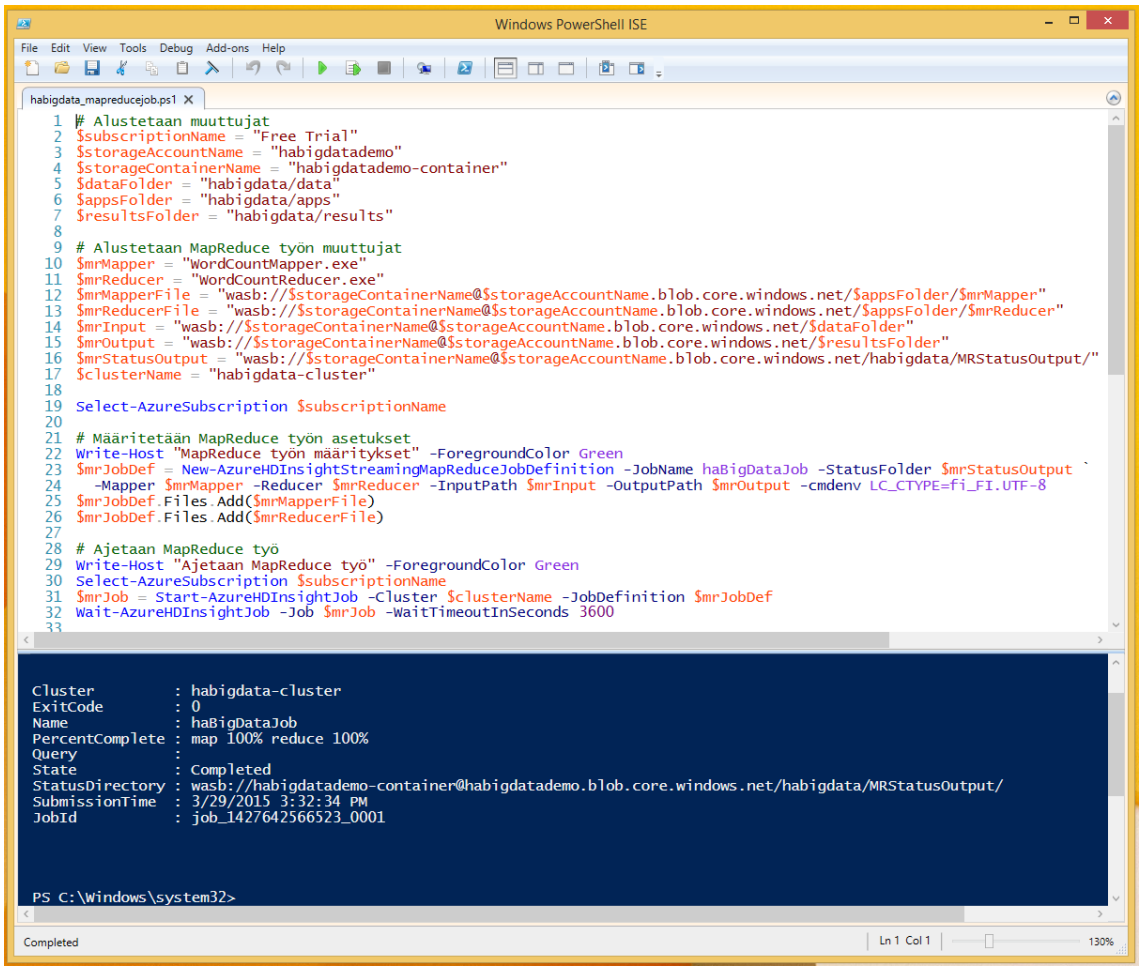
```
1  # Alustetaan muuttujat
2  $subscriptionName = "Free Trial"
3  $storageAccountName = "habigdatademo"
4  $storageContainerName = "habigdatademo-container"
5  $localAppsFolder = "C:\Users\Heikki\SkyDrive\Opinnaytetyo\demo\csharp-demo\Apps"
6  $destAppsFolder = "habigdata/apps"
7
8  # Tietovarasto
9  $storageAccountKey = (Get-AzureStorageKey -StorageAccountName $storageAccountName).Primary
10 Set-AzureSubscription -SubscriptionName $subscriptionName -CurrentStorageAccount $storageAccountName
11 $destContext = New-AzureStorageContext -StorageAccountName $storageAccountName -StorageAccountKey $storageAccountKey
12
13 # Siirretään sovellustiedostot paikalliselta levyltä Azure-palveluun
14 $appFiles = Get-ChildItem $localAppsFolder
15 foreach($file in $appFiles) {
16     $fileName = "$localAppsFolder\$file"
17     $blobName = "$destAppsFolder/$file"
18     Write-Host "Copying $fileName to $blobName"
19     Set-AzureStorageBlobContent -File $fileName -Container $storageContainerName -Blob $blobName -Context $destContext -Force
20 }
21 Write-Host "Kaikki sovellustiedostot hakemistosta $localAppsFolder ladattu $storageContainerName hakemistoon"
22
```

Kuva 18. Algoritmien siirtäminen Microsoft Azure -palveluun.

Riveillä 9–11 asetetaan kohteeksi oikea tiedostojärjestelmä, ja riveillä 14–20 siirretään kaikki tiedostot paikallisesta kehitysympäristöstä pilvipalvelun tiedostojärjestelmään.

Datan analysointi (MapReduce-työ)

Seuraavassa vaiheessa suoritettiin MapReduce-työ pilvipalvelussa. Tarpeelliset komennot nähdään kuvassa 19. MapReduce-työn parametrit määritellään komentotiedoston riveillä 23–26. New-AzureHDInsightStreamingMapReduceJobDefinition-komennolla määritellään niin sanottu Hadoop streaming -työ. Tämä tarkoittaa sitä, että Hadoop huolehtii datan lukemisesta hakemistosta tai tiedostosta, joka on määritelty parametrilla -InputPath ja kirjoittaa tulokset hakemistoon, joka on määritelty parametrilla -OutputPath.



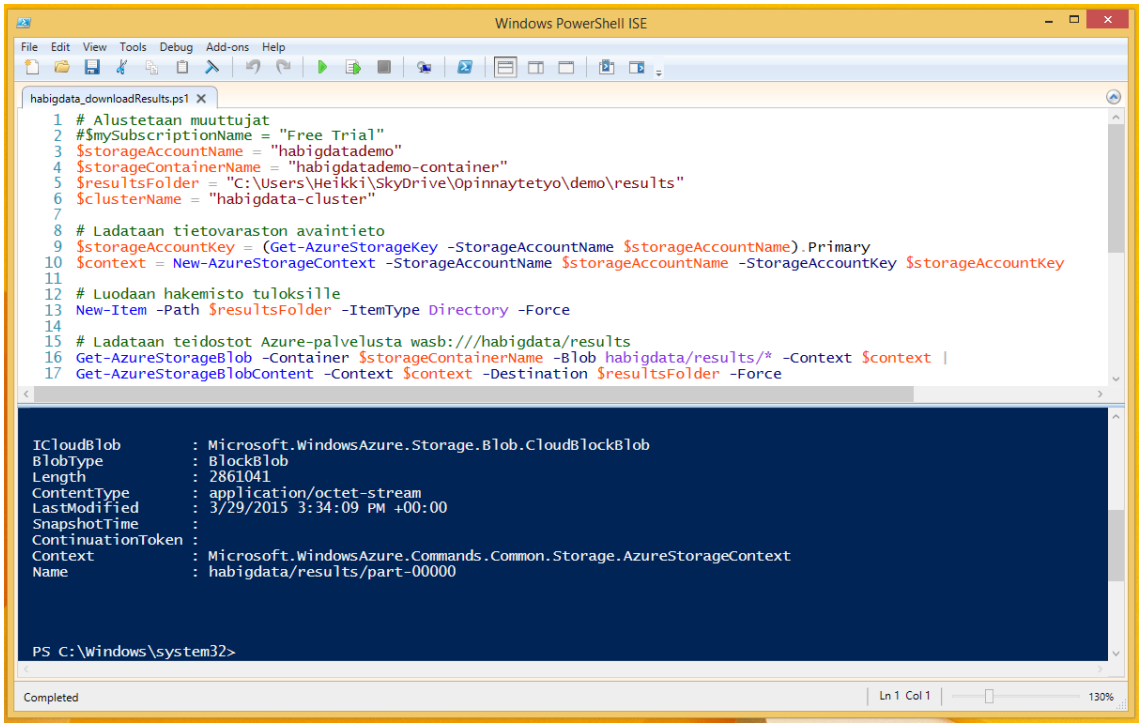
Kuva 19. Komennot MapReduce-työn ajamiseen Microsoft Azure -palvelussa.

Tässä ”streaming”-muotoisessa työssä data ikään kuin virtaa lähdetiedostoista Map- ja Reduce-ohjelmien kautta tulostiedostoon. Varsinainen MapReduce-työn suorittaminen aloitetaan rivin 31 komennolla Start-AzureHDInsightJob.

Powershell-työkalun alareunassa olevassa tulosikkunassa nähdään työn suorituksen lopputulos. Jos työ onnistui, ikkunassa on ExitCode-muuttujan arvona 0.

Tuloksien visualisointi

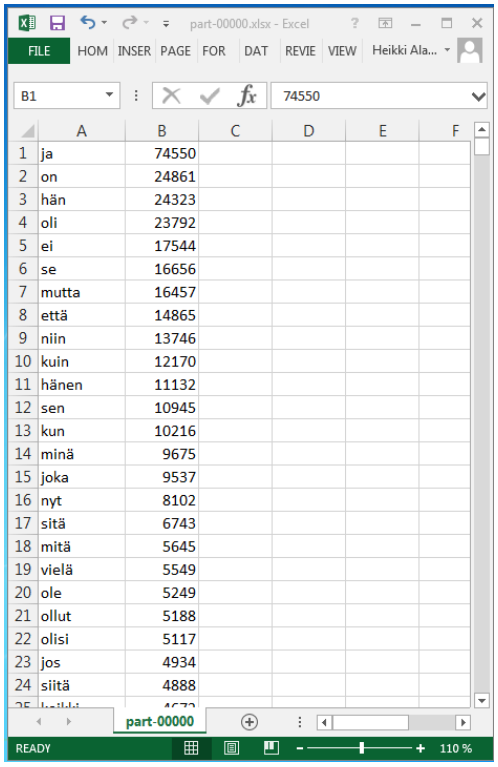
MapReduce-työn suorittamisen jälkeen siirrettiin tulostiedostot Azure-palvelusta paikalliselle levylle. Komentotiedosto tiedostojen siirtämiseen on esitetty kuvassa 20. Komentotiedostossa luodaan uusi hakemisto tulostiedostoille paikalliselle levylle. Hakemiston luonnin jälkeen tiedostot Azure-palvelusta siirretään sinne. Parametrilla -Force pakotetaan hakemistossa jo olemassa olevien samalla nimellä olevien tiedostojen päällekirjoitus.



Kuva 20. Komennot tulostiedostojen siirtämiseen Microsoft Azure -palvelusta.

Tulostiedosto on tekstimuotoinen tiedosto, jossa sanat ja niiden lukumäärä ovat aakkosjärjestyksessä erotettuna sarkaimella. Tuloksien käsittelyyn käytettiin Microsoft Excel -ohjelmaa. Tiedostoa avatessa oli huomioitava, että merkistöksi tuli valittua UTF-8, muuten tekstissä olevat ä- ja ö-merkit eivät näkyneet Excel-ohjelmassa oikein. Lisäksi ensimmäinen sarake tuli muotoilla tekstiksi.

Tulostiedosto lajiteltiin aluksi siten, että saatiin esille lähdeaineistossa eniten käytetyt sanat. Lajitellun tiedoston alku on esitetty kuvassa 21.

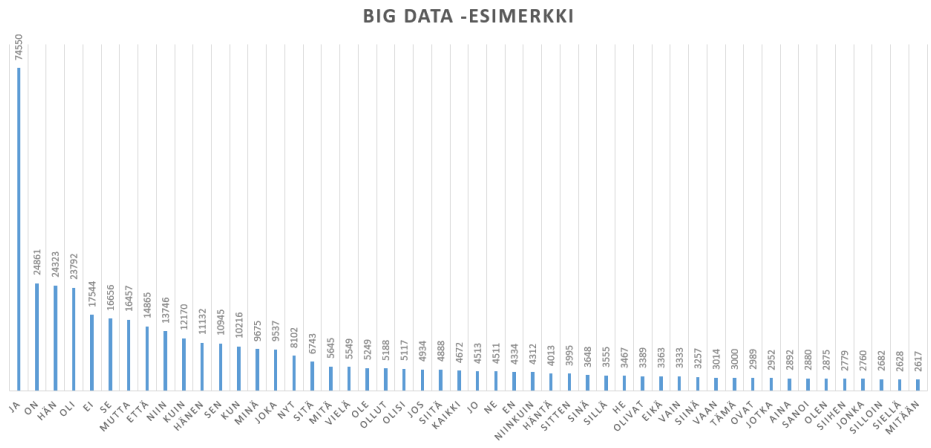


	A	B
1	ja	74550
2	on	24861
3	hän	24323
4	oli	23792
5	ei	17544
6	se	16656
7	mutta	16457
8	että	14865
9	niin	13746
10	kuin	12170
11	hänen	11132
12	sen	10945
13	kun	10216
14	minä	9675
15	joka	9537
16	nyt	8102
17	sitä	6743
18	mitä	5645
19	vielä	5549
20	ole	5249
21	ollut	5188
22	olisi	5117
23	jos	4934
24	siitä	4888

Kuva 21. Lajiteltu tulostiedosto.

Tulostiedostossa on eri sanoja yli 207 000, ja niistä 75 % esiintyy lähdeaineistossa vain kerran tai kaksi. Jotta tulokset pystyttäisiin visualisoimaan järkevästi, otetaan esityksiin vain osa tuloksista. Excel-ohjelma ja erilaiset internetpalvelut tarjoavat useita vaihtoehtoja esittää tulokset graafisesti. Seuraavassa esitellään kolme erilaista esimerkkiä.

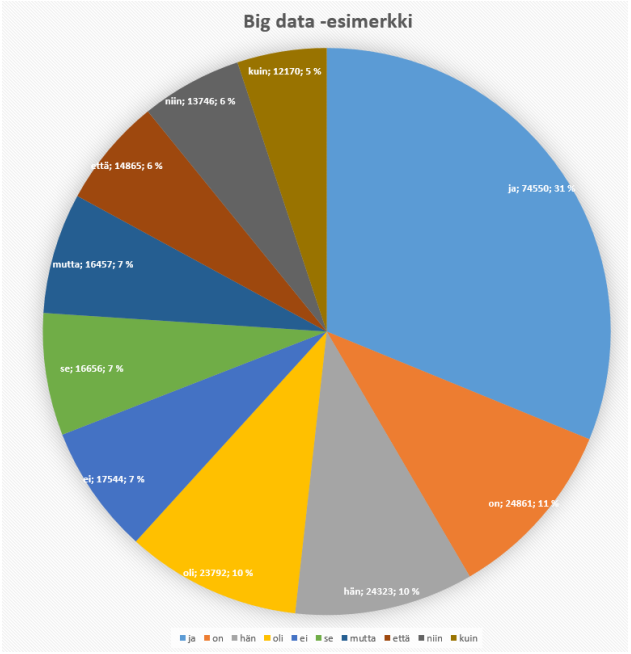
Kuvan 22 esimerkissä käytetään pylväskaaviota, jossa on x-akselilla esitettynä 50 eniten käytettyä sanaa.



Kuva 22. 50 eniten käytettyä sanaa.

Kaaviosta voidaan helposti nähdä sanojen esiintymistiheys, mutta ongelmana voi olla y-akselin skaalaus, kun yksi arvoista on muita huomattavasti suurempi.

Toisessa esimerkissä käytetään ympyräkaaviota esittämään 10 eniten esiintyneen sanan osuuksia kaikista lähdeaineiston sanoista. Esimerkki on kuvassa 23.



Kuva 23. 10 eniten käytettyä sanaa.

Ympyräkaaviosta nähdään nopeasti kuvattavien asioiden osuudet ja pystytään vertailemaan niitä toisiinsa. Lisäksi kaaviossa voidaan esittää tarkat prosenttiosuudet. Kovin montaa arvoa kaavioon ei kuitenkaan voi laittaa, muuten siitä tulee sekava.

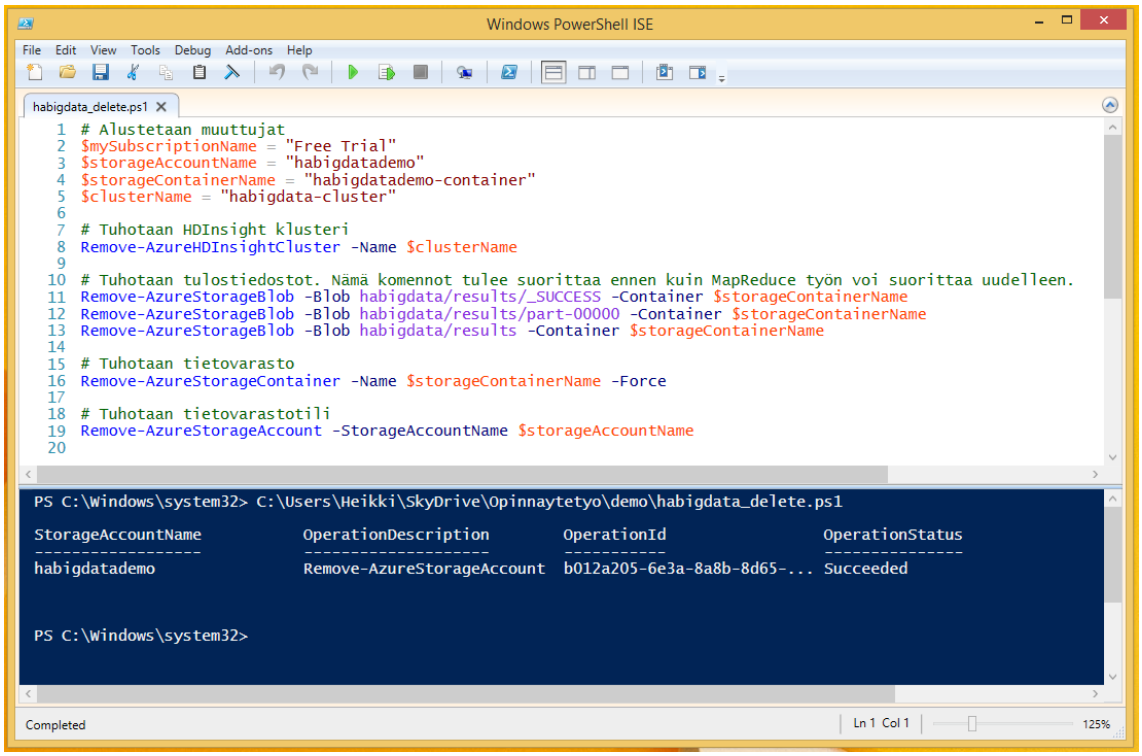
Lisäksi voidaan käyttää internetistä löytyviä palveluja [31], joilla tulokset voidaan esittää sanapilvenä. Esimerkki sanapilvestä on esitetty kuvassa 24.

Kuva 24. Sanapilvi.

Sanapilvestä voidaan saada nopea käsitys muutamasta tärkeimmästä asiasta tai tässä tapauksessa käytetyimmistä sanoista, mutta tarkkaa sanojen välisten suuruussuhteen eroa on vaikea nähdä.

HDInsight-klusterin ja tietovaraston poistaminen

Koska Microsoft Azure -palvelu on maksullinen ja tilauksen hinta pohjautuu palvelun kokoonpanoon ja käyttömäärään, siitä kannattaa poistaa kaikki ne palvelut, joita ei enää välttämättä tarvita.



Kuva 25. Komennot palvelujen poistamiseksi.

Demonstraatiossa tehtävän tulokset siirrettiin omalle paikalliselle levylle, joten kaikki luodut palvelut voitiin poistaa pilvipalvelusta. Komennot palvelujen poistoon on esitetty kuvassa 25. Palvelut poistetaan päinvastaisessa järjestyksessä kuin palvelut luotiin.

8 Yhteenveto

Insinöörityössä selvitettiin, mitä big data on käytännössä ja miten datasta saadaan aikaan tuloksia, joiden avulla voidaan parantaa liiketoimintaa tai luoda uutta liiketoimintaa. Lisäksi tutkittiin, millainen vaikutus big data -ilmiöllä on tavallisen ihmisen elämään ja millaisia mahdollisuuksia datan hyödyntäminen antaa yrityksille. Työssä käytiin myös läpi yleistä big datan käsittelyarkkitehtuuria ja erityisesti monessa ratkaisussa keskeistä komponenttia nimeltä Hadoop. Insinöörityössä määriteltiin prosessi datan käyttöönottoon ja havainnollistettiin sitä käytännössä Microsoft Azure HDInsight -palvelua käyttämällä.

Heti selvitystyön alussa ilmeni, että termiä big data voidaan käyttää monella eri tavalla. Koska termillä ei ole virallista yleispätevää määritelmää, sitä voidaan käyttää erilaisissa tilanteissa eri tavoilla. Lisäksi big data on noussut viime vuosina ilmiöksi, jonka uskotaan ratkaisevan kaikki mahdolliset ongelmat ainakin myyntimiesten puheissa. Toisessa äärilaidassa big data -ilmiötä on kritisoitu hyvinkin rankasti. Kritisoijat ovat sanoneet, että big datassa ei ole mitään uutta tai ainakaan mitään vallankumouksellista. Selvitystyössä havaittiin datan voivan olla uhka yksityisyydelle. Puutteellisesti anonymisoitu data voidaan avata esimerkiksi yhdistelemällä eri data-aineistoja. Tämä aiheuttanee tulevaisuudessa tarkempia säännöksiä datan julkaisulle.

Käyttöönottoprosessin luomisen ja demonstraation tekemisen yhteydessä havaittiin, ettei datan analysoinnin käyttöönotto ole ilman asiantuntijan olemassa oloa mahdollista. Prosessiin tarvitaan mukaan toimija, joka pystyy muuntamaan liiketoimintaongelman ohjelmoitavaksi algoritmiksi, käyttämään tarvittavia pilvipalveluja, visualisoimaan saadut tulokset ja kommunikoimaan ne. Tältä henkilöltä tai ryhmältä vaaditaan tietojenkäsittelyn, ohjelmoinnin, matematiikan ja tilastotieteen osaamisen lisäksi liiketoiminta-osaamista ja osaamista kyseessä olevalta toimialalta.

Demonstraatioympäristön Microsoft Azure HDInsight -palvelu pohjautuu paljon käytettyyn Apache Hadoop -ohjelmistokehykseen, kuten myös monet sen kilpailijat. Insinöörityötä voisi jatkaa tutkimalla, onko olemassa varteenotettavia vaihtoehtoja Hadoop-tekniikalle. Lisäksi Hadoop-ympäristöstäkin jäi demonstraation ulkopuolelle kaksi mielenkiintoista komponenttia, eli Pig ja Hive. Niitä käyttämällä demonstraatio voisi olla kattavampi.

Lähteet

1 Gartner's 2014 Hype Cycle for Emerging Technologies Maps the Journey to Digital Business. 2014. Verkkodokumentti. Gartner. <http://www.gartner.com/newsroom/id/2819918>. Päivitetty 11.8.2014. Luettu 1.2.2015.

2 Salo, Immo. 2014. Big data & pilvipalvelut. Jyväskylä: Docendo.

3 Big Data. 2014. Verkkodokumentti. Ivorio. <http://www.slideshare.net/ivoriofinland/big-data-esitys-joulukuu-2014>. Päivitetty 18.12.2014. Luettu 5.3.2015.

4 Laney, Doug. 2001. 3D Data Management Controlling Data Volume Velocity and Variety. Verkkodokumentti. Meta Group. < http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf >. Päivitetty 6.2.2001. Luettu 5.2.2015.

5 Salo, Immo. 2013. Big data tiedon vallankumous. Jyväskylä: Docendo.

6 Russom, Philip. 2011. Big Data Analytics. TDWI.

7 Sigular, Svetlana. 2013. Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. Verkkodokumentti. Forbes. <http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/2/>. Päivitetty 27.3.2013. Luettu 5.2.2015.

8 2 More Big Data V's — Value And Veracity. 2014. Verkkodokumentti. SAP. <http://blogs.sap.com/innovation/big-data/2-more-big-data-vs-value-and-veracity-01242817>. Päivitetty 23.1.2014. Luettu 26.3.2015.

9 Euroopan parlamentin ja neuvoston direktiivi 2003/98/EY. Verkkodokumentti. Euroopan unionin virallinen lehti. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:FI:PDF>. Päivitetty 17.11.2003. Luettu 8.3.2015.

10 de Montjoye, Yves-Alexandre, Radaelli, Laura, Singh, Vivek Kumar, Pentland, Alex. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. Science. 1/30/2015, Vol. 347 Issue 6221, s. 536–539.

11 Kaupan asiakasrekisterit avattiin botulismitapausten estämiseksi. 2011. Verkkodokumentti. Yle. <http://yle.fi/uutiset/kaupan_asiakasrekisterit_avattiin_botulismitapausten_estamiseksi/2988307>. Päivitetty 30.10.2011. Luettu 6.2.2015.

12 Seurataanko Plussa-ostoksia? – Keskolla selitys. 2011. Verkkodokumentti. Uusi Suomi. <http://www.uusisuomi.fi/kotimaa/117339-seurataanko-plussa-ostoksia-%E2%80%93%C2%A0keskolla-selitys>. Päivitetty 31.10.2011. Luettu 6.2.2015.

13 Henkilötietolaki 22.4.1999/523.

14 Big datan hyödyntäminen. Verkkodokumentti. Liikenne- ja viestintäministeriö. <http://www.lvm.fi/julkaisu/4417803/big-datan-hyodyntaminen>. Luettu 27.3.2015.

15 Tietosuojavaltuutetun lausunto liikenne- ja viestintäministeriön asettaman työryhmän valmistelemasta kansallisen big data -strategian luonnoksesta. Verkkodokumentti. Tietosuojavaltuutetun toimisto. <http://www.tietosuoja.fi/fi/index/ajankohtaista/tiedotteet/2014/06/tietosuojavaltuutetunlausuntoliikenne-javiestintaministerionasettamantyoryhmanvalmistelemastakansallisenbigdata-strategianluonnoksesta.html>. Päivitetty 27.6.2014. Luettu 27.3.2015

16 Balancion. 2014. Verkkodokumentti. Balancion. <http://www.balancion.com>. Luettu 1.3.2015.

17 Balancion käyttöehdot. 2014. Verkkodokumentti. Balancion. <http://www.balancion.com/kayttoehdot>. Päivitetty 1.4.2014. Luettu 1.3.2015.

18 Mayer-Schönberger, Viktor, Cukier, Kenneth. 2013. Big Data. A revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt Publishing Company.

19 Google books Ngram Viewer. 2013. Verkkodokumentti. Google. <https://books.google.com/ngrams>. Luettu 25.2.2015.

20 Helmiä kalastamassa - Avauksia tietämyksen hallintaan. Verkkodokumentti. Eduskunnan kanslian julkaisu 1/2001. <http://www.eduskunta.fi/triphome/bin/thw.cgi/thw.cgi/trip/?${APPL}=erekj&${BASE}=erekj&${THWIDS}=0.8/1427280188_151736&${TRIPPIFE}=PDF.pdf>. Päivitetty 2001. Luettu 25.3.2015.

21 Mohanty, Soumendra, Jagadeesh, Madhu, Srivatsa, Harsha. 2013. Big Data Imperatives. New York: APress.

22 Krum, Randy. 2014. Cool infographics : effective communication with data visualization and design. Indianapolis: Wiley.

23 Saarelainen, Ari. 2015. Ison datan kalastajat. TiVi. 1/2015, s. 26–35.

24 Sawant, Nitin, Shah, Himanshu. 2013. Big Data Application Architecture Q&A: A Problem - Solution Approach. New York: APress.

25 Hadoop. Verkkodokumentti. The Apache Software Foundation. <http://hadoop.apache.org/>. Päivitetty 27.3.2015. Luettu 2.4.2015.

26 White, Tom. 2011. Hadoop. The Definitive Guide. Sebastobol: O'Reilly Media.

27 Prosessien mallintaminen osana toiminnan kehittämistä. Verkkodokumentti. Tampereen teknillinen yliopisto. <http://dspace.cc.tut.fi/dpub/bitstream/handle/123456789/6825/prosessien_mallintaminen.pdf>. Päivitetty 13.12.2010. Luettu 3.4.2015.

28 Chauhan, Avkash, Fontama, Valentine, Hart, Michele, Hyong Tok, Wee, Woody, Buck. 2014. Introducing Microsoft Azure HDInsight Technical Overview. Redmond: Microsoft Press.

29 Free ebooks - Project Gutenberg. Verkkodokumentti. Project Gutenberg. < http://www.gutenberg.org/wiki/Main_Page >. Luettu 8.3.2015.

30 How to install and configure Azure PowerShell. Verkkodokumentti. Microsoft. <http://azure.microsoft.com/en-us/documentation/articles/powershell-install-configure/>. Luettu 12.3.2015.

31 WordItOut. 2015. Verkkodokumentti. Enideo. < http://worditout.com/>. Luettu 5.4.2015.

Hype Cycle for Emerging Technologies, 2014